The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition

Marcus Thomas Pearce

Doctor of Philosophy

Department of Computing
City University, London

December 2005

ABSTRACT

The prevalent approach to developing cognitive models of music perception and composition is to construct systems of symbolic rules and constraints on the basis of extensive music-theoretic and music-analytic knowledge. The thesis proposed in this dissertation is that statistical models which acquire knowledge through the induction of regularities in corpora of existing music can, if examined with appropriate methodologies, provide significant insights into the cognitive processing involved in music perception and composition. This claim is examined in three stages. First, a number of statistical modelling techniques drawn from the fields of data compression, statistical language modelling and machine learning are subjected to empirical evaluation in the context of sequential prediction of pitch structure in unseen melodies. This investigation results in a collection of modelling strategies which together yield significant performance improvements over existing methods. In the second stage, these statistical systems are used to examine observed patterns of expectation collected in previous psychological research on melody perception. In contrast to previous accounts of this data, the results demonstrate that these patterns of expectation can be accounted for in terms of the induction of statistical regularities acquired through exposure to music. In the final stage of the present research, the statistical systems developed in the first stage are used to examine the intrinsic computational demands of the task of composing a stylistically successful melody. The results suggest that the systems lack the degree of expressive power needed to consistently meet the demands of the task. In contrast to previous research, however, the methodological framework developed for the evaluation of computational models of composition enables a detailed empirical examination and comparison of such models which facilitates the identification and resolution of their weaknesses.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisors Geraint Wiggins, Darrell Conklin and Eduardo Alonso for their guidance and support in both academic and administrative matters during the course of the research reported in this dissertation. I am also indebted to my friends and colleagues at City University and elsewhere for providing a stimulating intellectual environment in which the present research was carried out. In particular, many thanks are due to Tak-Shing Chan, David Meredith, Christopher Pearce, Alison Pease, Christophe Rhodes and Kerry Robinson for their detailed comments on earlier drafts of material appearing in this dissertation. This dissertation also benefited enormously from the careful reading of my examiners, Petri Toiviainen and Artur d'Avila Garcez. In addition, Alan Pickering provided useful advice on statistical methodology. I would also like to acknowledge the support of Andrew Pearce in the music department at City University, John Drever in the music department at Goldsmiths College as well as Aaron Williamon and Sam Thompson at the Royal College of Music who went out of their way to help me in recruiting judges for the experiments reported in Chapter 9 and also Darrell Conklin for providing the experimental data used in §8.7. Finally, the research presented in this dissertation would not have been possible without the financial support of City University, who provided funds for equipment and conference expenses, and the Engineering and Physical Sciences Research Council (EPSRC) who supported my doctoral training via studentship number 00303840.

* * *

Marcus T. Pearce
7 December 2005

CONTENTS

List of Tables

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 The Problem Domain and Approach

The research presented in this dissertation is concerned with modelling cognitive processes in the perception and composition of melodies. The particular computational problem studied is one of *sequence prediction*: given an ordered sequence of discrete events, the goal is to predict the identity of the next event (Dietterich & Michalski, 1986; Sun & Giles, 2001). In general, the prediction problem is non-deterministic since in most stylistic traditions an incomplete melody may have a number of plausible continuations.

Broadly speaking, we adopt an empiricist approach to solving the problem, in which the function governing the identity of an event in a melodic sequence is learnt through experience of existing melodies. In psychology, learning is usually defined as "the process by which long-lasting changes occur in behavioural potential as a result of experience" (Anderson, 2000, p. 4). Expanding on this definition, research in machine learning specifies a well-posed learning problem as one in which the source of experience is identified and the changes in behavioural potential are quantified as changes in a performance measure on a specified set of tasks:

> A computer program is said to *learn* from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.
>
> (Mitchell, 1997, p. 2)

As stated above, the task T is one of non-deterministic sequence prediction in which, given a sequence $s_i, s_{i+1}, \ldots, s_j$, the goal is to predict s_{j+1}. Having predicted s_{j+1}, the learner is shown s_{j+1} and challenged to predict s_{j+2} and so on. This differs from the classification problems typically studied in machine learning where the goal is to learn the function mapping examples from the target domain onto a discrete set of class labels (Sun & Giles, 2001). The performance measure P is the performance of the trained model in predicting unseen melodies, operationalised in terms of the average surprisal induced in the model by each unseen event. Finally, the source of experience E consists of melodies drawn from existing musical repertoires.

Machine learning algorithms differ along a number of dimensions. For example, it is common to distinguish between *inductive* learning and *analytical* learning. While the former involves statistical inference on the basis of existing data to find hypotheses that are consistent with the data, the latter involves deductive inference from a logical domain theory to find hypotheses that are consistent with this theory. Analytical learners can learn from scarce data but require the existence of significant *a priori* domain knowledge. Inductive learners, on the other hand, require little prior knowledge of the domain but require extensive data from which to learn. Furthermore, in order to generalise to novel domain examples, inductive learning algorithms require an inductive bias: a set of assumptions about the target hypothesis, which serve to justify its inductive inferences as deductive inferences (Mitchell, 1997). Inductive learning algorithms are also commonly classified according to whether they learn in a *supervised* or *unsupervised* manner. Supervised learning algorithms require feedback during learning as to the correct output corresponding to any given input, while unsupervised learners require no such feedback. The selection of an appropriate kind of machine learning algorithm (supervised or unsupervised; inductive or analytical) is heavily task dependent, depending on the relative availability of large corpora of training data, extensive domain theories and target outputs.

In the present research, an unsupervised, inductive learning approach is followed, which makes minimal *a priori* assumptions about the sequential structure of melodies. The particular brand of inductive learning model examined may be categorised within the class of *finite context* or n-gram models. Introduced fully in §3.2 and §6.2.1, these models represent knowledge about a target domain of sequences in terms of an estimated probability distribution governing the identity of an event given a context of preceding events in the sequence. The length of the context is referred to as the *order* of the model. As discussed in §3.2, these models are intrinsically weak in terms of the structural descrip-

tions they assign to sequences of events (although this weakness is orthogonal to their stochastic nature). However, in contrast to more powerful modelling approaches, finite context models lend themselves to an unsupervised learning approach in which the model acquires its knowledge of sequential structure in the target domain exclusively through exposure to existing event sequences drawn from that domain. Finally, the research presented in this dissertation emphasises the problem of accurately estimating event probabilities from trained models (and examining these models in the context of music cognition) rather than comparing the performance of different learning algorithms.

1.2 Motivations: Cognition, Computation and Analysis

Existing cognitive models of music perception typically consist of systems of symbolic rules and constraints constructed by hand on the basis of extensive (style specific) music-theoretic knowledge (*e.g.,* Deutsch & Feroe, 1981; Lerdahl & Jackendoff, 1983; Narmour, 1990; Temperley, 2001).[1] The same may be said of research on cognitive processes in music composition (*e.g.,* Baroni, 1999; Johnson-Laird, 1991) although this area of research has received far less attention than the perception of music. When inductive statistical models of observed phenomena in music perception have been examined (see §3.6), they have typically been limited to fixed, low order models of a small number of simple representational dimensions of music (Eerola, 2004b; Krumhansl, 1990; Krumhansl *et al.*, 1999; Oram & Cuddy, 1995; Vos & Troost, 1989).

Within the field of Artificial Intelligence (AI), sophisticated statistical learning models which operate over rich representations of musical structure have been developed (see §3.4) and used for a number of tasks including the prediction of music (Conklin & Witten, 1995), classification of music (Westhead & Smaill, 1993) and stylistic analysis (Ponsford *et al.*, 1999). In particular, the *multiple viewpoints framework* (Conklin & Witten, 1995) extends the use of finite context modelling techniques to domains, such as music, where events have an internal structure and are richly representable in languages other than the basic event language (see §5.2.3). However, this body of research has not examined the capacity of such models to account for observed phenomena in music perception. Furthermore, while the models developed have been used to generate music, the objective has been to verify the music analytic principles involved in their construction (Conklin & Witten, 1995; Ponsford *et al.*, 1999)

[1] The theory of Lerdahl & Jackendoff (1983) is summarised in §3.3 and that of Narmour (1990) in §8.2.2.

or to examine their utility as tools for composers and performers (Assayag *et al.*, 1999; Lartillot *et al.*, 2001) and not specifically to model cognitive processes in music composition.

The motivation behind the research presented in this dissertation is to address the observed gulf between the development of sophisticated statistical models of musical structure in AI research and their application to the understanding of cognitive processing in music perception and composition. It is pertinent to ask, however, whether there is any reason to believe that addressing this issue will afford any advantages over and above existing approaches in the study of music cognition. As noted above, the dominant theories of music cognition consist of hand constructed systems of symbolic rules and constraints derived from extensive and specialised music-analytic knowledge. Without a doubt, such theories have made significant contributions to the understanding of music cognition in terms of explicit accounts of the structures potentially afforded by the perceptual environment. However, as noted by West *et al.* (1985) and suggested by a small number of empirical studies (Boltz & Jones, 1986; Cook, 1987), these theoretical accounts may significantly overestimate the perceptual and cognitive capacities of even musically trained listeners. Furthermore, as noted by Cross (1998a), they are typically accompanied by claims of universal applicability and exhibit a degree of inflexibility which are incommensurate with the small number of empirical psychological studies of music perception in cross-cultural settings (Castellano *et al.*, 1984; Eerola, 2004b; Stobart & Cross, 2000).

From a methodological perspective, Cook (1994) charges the prevalent approaches in music cognition with *theorism*, the implicit premise that people perceive music in terms of music-theoretic structures which were, in fact, developed for pedagogical purposes. In considering this tension between music theory and music psychology, Gjerdingen (1999a, pp. 168–169) encourages the use of machine learning models to develop "theories of music perception that replace the calculus of musical atoms with an emphasis on experience, training and attention." In summary, the application of sophisticated techniques for knowledge acquisition and deployment to the development of data-driven models of music cognition offers the opportunity of addressing the theory-driven biases, inflexibility and cross-cultural limitations of current approaches to the modelling of music cognition.[2]

[2]As discussed in §2.6, the machine learning approach also affords other related methodological advantages.

1.3 Thesis Statement

The thesis proposed in this dissertation is that statistical models which acquire knowledge through induction of regularities in corpora of existing music can, if examined with appropriate methodologies, provide significant insights into the cognitive processing involved in music perception and composition. In particular, the present research seeks answers to the following specific questions:

1. Which computational techniques yield statistical models of melodic structure that exhibit the best performance in predicting unseen melodies?
2. Can these models account for empirically observed patterns of expectation exhibited by humans listening to melodies?
3. Can these models account for the cognitive processing involved in composing a stylistically successful melody?

In pursuing answers to each of these questions, it is necessary to decide upon a methodological approach which is capable of producing empirical results pertinent to answering the question. Where appropriate methodologies exist in relevant fields of research, they have been adopted; in addition, it is within the scope of the present research to adapt or elaborate existing methodologies in order to yield objective answers to the research questions (see, for example, Chapter 9). In the case of Question 1, the techniques examined as well as the methodologies used to evaluate these techniques are drawn from research in the fields of Artificial Intelligence and Computer Science. However, Questions 2 and 3 explicitly introduce the goal of understanding cognitive processes which in turn implies different criteria and methodological approaches for evaluating the computational models (see §2.4). Since our current understanding of statistical processes in music perception and, especially, composition is relatively undeveloped, the present research follows common practice in cognitive-scientific research in adopting a computational level approach (see §2.4). Specifically, the focus is placed on developing our understanding of the intrinsic nature and computational demands of the tasks of perceiving melodic structure and composing a melody in terms of constraints placed on the expressive power and representational dimensions of the cognitive systems involved.

1.4 Research Objectives and Scope

Given the motivating factors discussed in §1.2 and the research questions stated in §1.3, the research presented in this dissertation adopts the following specific

objectives:

1. to conduct an empirical examination of a range of modelling techniques in order to develop powerful statistical models of musical structure which have the potential to account for aspects of the cognitive processing of music;

2. to apply the best performing of these models in an examination of specific hypotheses regarding cognitive processing in music perception and composition;

3. to investigate and adopt appropriate existing methodologies, adapting and elaborating them where necessary, for the empirical evaluation of these hypotheses.

In order to reduce the complexity of the task of achieving these objectives, the scope of the research presented in this dissertation was constrained in several ways. First, the present research is limited to modelling monophonic music and the corroboration of the results with homophonic or polyphonic music remains a topic for future research (see §4.2).[3] Second, the focus is placed firmly on modelling pitch structure, although the influences of tonal, rhythmic, metric and phrase structure on pitch structure are taken into consideration (see §5.4). This decision may be justified in part by noting that pitch is generally the most complex dimension of the musical genres considered in the present research (see §4.3). Third, a symbolic representation of the musical surface is assumed in which a melody consists of a sequence of discrete events which, in turn, are composed of a finite number of discrete features (see §5.1). This decision may be justified by noting that many aspects of music theory, perception and composition operate on musical phenomena defined at this level (Balzano, 1986b; Bharucha, 1991; Krumhansl, 1990; Lerdahl, 1988a). Fourth, several complex features, such as tonal centres or phrase boundaries, are taken directly from the score (see §5.3). It is assumed that the determination of these features in a given task such as melody perception may be regarded as a subcomponent of the overall problem to be solved independently from the present modelling concerns.

In addition to these constraints imposed on the nature and representation of the objects of study, some limitations were placed on the modelling techniques used. In particular, the present research examines the minimal requirements

[3] A piece of music is monophonic if it is written for a single voice, homophonic if it is written for multiple voices all of which move in the same rhythm and polyphonic if it is written for multiple voices each exhibiting independent rhythmic movement.

placed on the cognitive processing of melodies through the exclusive use of finite context models (see §3.2). If these relatively weak grammars prove insufficient to meet the demands of a given task, it remains for future research to examine the capacity of more powerful grammars on that task. This decision may be justified by invoking the principle of Ockham's razor: we prefer simpler models which make fewer assumptions until the limited capacities of such models prove inadequate in accounting for empirically observed phenomena.

1.5 Original Contributions

In §2.3, a distinction is made between three different branches of AI each with its own motivations, goals and methodologies: basic AI; cognitive science; and applied AI. The present research makes direct contributions in the fields of basic AI and, especially, cognitive science and indirectly contributes to the field of applied AI.

The goal of basic AI is to examine computational techniques which have the potential for simulating intelligent behaviour. Chapters 6 and 7 present an examination of the potential of a range of computational modelling techniques to simulate intelligent behaviour in the context of sequence learning and prediction. The techniques examined and the methodologies used to evaluate these techniques are drawn from the fields of data compression, statistical language modelling and machine learning. In particular, Chapter 6 examines a number of strategies for deriving improved predictions from trained finite context models of melodic pitch structure, whilst Chapter 7 introduces a new technique based on a weighted geometric mean for combining the predictions of multiple models trained on different representations of the musical surface. In empirically identifying a number of techniques which consistently improve the performance of finite context models of melodic music, the present research contributes to our basic understanding of computational models of intelligent behaviour in the induction and prediction of musical structure.

Another contribution made in the present research is to use a feature selection algorithm to construct multiple viewpoint systems (see 5.2.3) on the basis of objective criteria rather than hand-crafting them on the basis of expert human knowledge as has been done in previous research (Conklin, 1990; Conklin & Witten, 1995). This allows the empirical examination of hypotheses regarding the degree to which different representational dimensions of a melody afford regularities which can be exploited by statistical models of melodic structure and in music cognition.

The goal of cognitive-scientific research is to further our understanding of human cognition using computational techniques. In Chapter 8, the statistical techniques developed in Chapters 6 and 7 are used to analyse existing behavioural data on melodic expectations. The results support the theory that expectations are generated by a cognitive system of unsupervised induction of statistical regularities in existing musical repertoires. This theory provides a functional account, in terms of underlying cognitive mechanisms, of existing theories of expectancy in melody (Narmour, 1990) and addresses the theory-driven biases associated with such knowledge-engineering theories (see §1.2). It also offers a more detailed and parsimonious model of the influences of the current musical context and prior musical experience on music perception.

In Chapter 9, computational constraints on melodic composition are examined by applying the statistical techniques developed in Chapters 6 and 7 to the task of generating stylistically successful melodies. In spite of efforts made to improve on the modelling strategies adopted in previous research, the results demonstrate that these simple grammars are largely incapable of meeting the intrinsic demands of the task. Given that the same models successfully accounted for empirically observed phenomena in music perception, this result is significant in the light of arguments made in previous research that similar grammars underlie the perception and composition of music (Baroni, 1999; Lerdahl, 1988a). In addition, the methodology developed to evaluate the computational systems constitutes a significant contribution to future research in the cognitive modelling of composition.

Finally, the goal of applied AI is to use existing AI techniques to develop applications for specific purposes in industry. While this is not a direct concern in the present research, the contributions made in terms of basic AI and cognitive science could be put to practical use in systems for computer-assisted composition (Ames, 1989; Assayag *et al.*, 1999; Hall & Smith, 1996), machine improvisation with human performers (Lartillot *et al.*, 2001; Rowe, 1992) and music information retrieval (Pickens *et al.*, 2003). Therefore, although these practical applications are not investigated in this dissertation, the research presented here constitutes an indirect contribution to such fields of applied AI.

1.6 Dissertation Outline

Background and Methodology

Chapter 2 contains a discussion of relevant epistemological and methodological issues concluding with an examination of the implications such issues raise

for the selection of appropriate methodologies for achieving the goals of the present research.

Chapter 3 presents the background on the modelling techniques used in the present research as well as a review of previous research which has applied them and related techniques to modelling music and music cognition.

Music Corpora and Representation

Chapter 4 contains a discussion of issues involved in the selection of data for computational modelling of music and presents the corpora of melodic music used in the present research.

Chapter 5 reviews several existing formal schemes for the representation of music and introduces the multiple viewpoint framework developed in the present research for the flexible representation and processing of a range of different kinds of melodic structure. The individual attribute types implemented are motivated in terms of previous research on music cognition and the computational modelling of music.

Statistical Modelling of Melodic Structure

Chapter 6 examines a number of techniques for improving the prediction performance of finite context models of pitch structure. These techniques, drawn primarily from research on statistical language modelling and data compression, are subjected to empirical evaluation on unseen melodies in a range of styles leading to significant improvements in prediction performance.

Chapter 7 introduces prediction within the context of multiple viewpoint frameworks. A new method for combining the predictions of different models is presented and empirical experiments demonstrate that it yields improvements in performance over existing techniques. A further experiment investigates the use of feature selection to derive multiple viewpoint systems with improved prediction performance.

Cognitive Processing of Melodic Structure

Chapter 8 presents the application of the statistical systems developed in the foregoing two chapters to the task of modelling expectancy in melody perception. In contrast to previous accounts, the results demonstrate that observed

patterns of melodic expectation can be accounted for in terms of the induction of statistical regularities acquired through exposure to music.

Chapter 9 describes the use of several multiple viewpoint systems developed in previous chapters to generate new chorale melodies in an examination of the intrinsic computational demands of composing a successful melody. The results demonstrate that none of the systems meet the demands of the task in spite of efforts made to improve upon previous research on music generation from statistical models. In contrast to previous approaches, however, the methodological framework developed for the evaluation of the computational systems enables a detailed and empirical examination and comparison of the systems leading to the identification and resolution of some of their salient weaknesses.

Summary and Conclusions

Chapter 10 includes a summary review of the research presented in this dissertation, a concise statement of the contributions and limitations of this research and a discussion of promising directions for developing the contributions and addressing the limitations in future research.

1.7 Publications

Parts of this dissertation are based on the following research papers which have been accepted for publication in journals and conference proceedings during the course of the present research. All of these papers were peer reviewed prior to publication.

Pearce, M. T., Conklin, D., & Wiggins, G. A. (2005). Methods for combining statistical models of music. In Wiil, U. K. (Ed.), *Computer Music Modelling and Retrieval*, (pp. 295–312). Heidelberg, Germany: Springer.

Pearce, M. T., Meredith, D., & Wiggins, G. A. (2002). Motivations and methodologies for automation of the compositional process. *Musicæ Scientiæ*, *6*(2), 119–147.

Pearce, M. T. & Wiggins, G. A. (2002). Aspects of a cognitive theory of creativity in musical composition. In *Proceedings of the ECAI'02 Workshop on Creative Systems*, (pp. 17–24). Lyon, France.

Pearce, M. T. & Wiggins, G. A. (2003). An empirical comparison of the performance of PPM variants on a prediction task with monophonic music. In *Proceedings of the AISB'03 Symposium on Artificial Intelligence and Creativity in Arts and Science*, (pp. 74–83). Brighton, UK: SSAISB.

Pearce, M. T. & Wiggins, G. A. (2004). Rethinking Gestalt influences on melodic expectancy. In Lipscomb, S. D., Ashley, R., Gjerdingen, R. O., & Webster, P. (Eds.), *Proceedings of the 8th International Conference of Music Perception and Cognition*, (pp. 367–371). Adelaide, Australia: Causal Productions.

Pearce, M. T. & Wiggins, G. A. (2004). Improved methods for statistical modelling of monophonic music. In *Journal of New Music Research*, *33*(4), 367–385.

Pearce, M. T. & Wiggins, G. A. (2006). Expectation in melody: The influence of context and learning. To appear in *Music Perception*.

CHAPTER 2

EPISTEMOLOGICAL AND METHODOLOGICAL FOUNDATIONS

2.1 Overview

The aim in this chapter is to define appropriate methodologies for achieving the objectives of the present research as specified in §1.4. Since an empirical scientific approach is adopted for the study of a phenomenon, music, which is traditionally studied in the arts and humanities, the first concern is to distinguish scientific from non-scientific methodologies (see §2.2). The current research examines music, specifically, from the point of view of Artificial Intelligence (AI) and in §2.3 three branches of AI are introduced, each of which has its own motivations and methodologies. The present research falls into the cognitive-scientific tradition of AI research and in §2.4, the dominant methodologies in cognitive science are reviewed. Given this general methodological background, §2.5 contains a discussion of methodological concerns which arise specifically in relation to the study of music from the perspective of science and AI. Finally, in §2.6 appropriate methodologies are defined for achieving the objectives of the present research based on the issues raised in the foregoing sections.

2.2 Speculative and Empirical Disciplines

Speculative disciplines are characterised by the use of deduction from definitions of concepts, self-evident principles and generally accepted propositions. Typically following a hermeneutic approach, "Their ultimate criterion of valid-

ity is whether they leave the reader with a feeling of conviction" (Berlyne, 1974, p. 2). Such fields as the aesthetics of music, music history and music criticism fall into this category. *Empirical* disciplines, on the other hand, are those which adopt experimental, scientific methodologies. It is important to be clear about the meaning of the term *science* since:

> A great deal of confusion has arisen from failure to realise that words like the French *science* and the German *Wissenschaft* (with their equivalents in other European languages) do not mean what the English word "science" means. A more accurate translation for them would be "scholarship".
>
> (Berlyne, 1974, p. 3)

Since we shall be adopting an empirical approach to the study of a phenomenon, music, which is traditionally examined from a speculative point of view, it will be helpful to preface this inquiry with a discussion of the epistemological status of scientific knowledge.

In *The Logic of Scientific Discovery*, Karl Popper (1959) developed an epistemological approach known as *methodological falsificationism* in an attempt to distinguish (systems of) propositions in the scientific disciplines from those of non-scientific fields. Popper rejected the verifiability criterion of logical positivism (the assertion that statements are meaningful only insofar as they are verifiable) on two grounds: first, it does not characterise the actual practice of scientific research; and second, it both excludes much that we consider fundamental to scientific inquiry (*e.g.,* the use of theoretical assumptions which may not be verifiable even in principle) and includes much that we consider non-scientific (*e.g.,* astrology). According to Popper, scientific statements must be embedded in a framework that will potentially allow them to be refuted:

> statements, or systems of statements, convey information about the empirical world only if they are capable of clashing with experience; or, more precisely, only if they can be *systematically tested*, that is to say, if they can be subjected . . . to tests which *might* result in their refutation.
>
> (Popper, 1959, pp. 313–314)

In logical terms, Popper's thesis stems from the fact that while an existential statement (*e.g.,* 'the book in front of me is rectangular') can be deduced from a universal statement (*e.g.,* 'all books are rectangular'), the reverse is not true. It

is impossible to verify a universal statement by looking for instances which confirm that statement (*e.g.,* by looking for rectangular books). We may only evaluate a universal statement by looking for empirical data supporting an existential statement that falsifies that statement (*e.g.,* by looking for non-rectangular books). According to Popper, a theory is only scientific if there exist existential statements which would refute the theory. The demarcation criterion also demands that a scientific theory must be stated clearly and precisely enough for it to be possible to decide whether or not any existential statement conflicts with the theory.

In methodological terms, falsificationism suggests that science does not consist of a search for truth but involves the construction of explanatory hypotheses and the design of experiments which may refute those hypotheses. A theory that goes unrefuted in the face of empirical testing is said to have been *corroborated*. Popper acknowledged that "scientific discovery is impossible without a faith in ideas which are of a purely speculative kind" (Popper, 1959, p. 25). However, he argued that the experiments designed to refute a scientific hypothesis must be empirical in nature in order for them to be intersubjectively tested. Therefore, the demarcation between scientific and non-scientific theories relies not on degree of formality or precision nor on weight of positive evidence but simply on whether empirical experiments which may refute those theories are proposed along with the hypotheses (see Gould, 1985, ch. 6, for an exposition of this thesis).

Although Popper remains to this day one of the most influential figures in scientific epistemology, he has received his fair share of criticism. In particular, several authors have argued that his account fails to accurately describe the actual progress of scientific research (Kuhn, 1962; Lakatos, 1970). Kuhn (1962) argued that in *normal science* researchers typically follow culturally defined paradigms unquestioningly. When such paradigms begin to fail, a crisis arises and gives rise to a scientific revolution which is caused not by rational or empirical but sociological and psychological factors: "... in Kuhn's view scientific revolution is irrational, a matter for mob psychology" (Lakatos, 1970, p. 91). It should be noted, however, that Kuhn's account is motivated more by descriptive concerns than the prescriptive concerns of Popper.

Imre Lakatos (1970), however, attempted to address Kuhn's criticisms of Popper's *naïve* falsificationism. In his own *sophisticated methodological falsificationism*, the basic unit of scientific achievement is not an isolated hypothesis but a *research programme* which he describes (at a mature stage of development) in terms of a theoretical and irrefutable *hard core* surrounded by a *protective*

belt of more flexible hypotheses each with their own problem solving machinery (Lakatos, 1970). The hard core of a programme is defined by its *negative heuristic*, which specifies which directions of research to avoid (those which may not refute the hard core), and its *positive heuristic*, which suggests fruitful research agendas for the reorganisation of the protective belt. The hard core is developed progressively as elements in the protective belt continue to go unrefuted.

Under this view, research programmes may be divided into those which are *progressive*, when they continue to predict novel facts as changes are continually made to the protective belt and hard core, or *degenerating*, when they lapse into constant revision to explain facts *post hoc*. Therefore, whole research programmes are not falsified by experimental refutation alone but only through substitution by a more progressive programme which not only explains the previous unrefuted content of the old programme and makes the same unrefuted predictions, but also predicts novel facts not accounted for by the old programme. Sophisticated methodological falsificationism seems to characterise well the actual progress of science (Lakatos, 1970) and "is an increasingly popular view of change in scientific theories" (Brown, 1989, p. 7).

2.3 Artificial Intelligence

Noting that it is possible to differentiate *natural science* (the study and understanding of natural phenomena) from *engineering science* (the study and understanding of practical techniques), Bundy (1990, p. 216) argues that there exist three branches of AI:

1. *basic AI*: an engineering science whose aim is to "explore computational techniques which have the potential for simulating intelligent behaviour";

2. *cognitive science* or *computational psychology*: a natural science whose aim is "to model human or animal intelligence using AI techniques";

3. *applied AI*: epistemologically speaking a branch of engineering "where we use existing AI for commercial techniques, military or industrial products, *i.e.,* to build products".

Since research in the different disciplines is guided by different motivations and aims, this taxonomy implies different "criteria for assessing research in each kind of AI. It suggests how to identify what constitutes an advance in the subject and it suggests what kind of methodology AI researchers might adopt" (Bundy,

1990, p. 219).[1] In accordance with this analysis, Wiggins & Smaill (2000) note that the motivations for applying AI techniques to the musical domain can be drawn out on a continuum between those concerned with understanding human musical abilities at one extreme (cognitive science) and those concerned with designing useful tools for musicians, composers and analysts at the other (applied AI).

2.4 Cognitive Science

The theoretical hard core in the overall research programme of cognitive science may be defined in terms of its negative and positive heuristics (see §2.2). The overriding negative heuristic is that purely behavioural or purely biological approaches to understanding cognition are unlikely to prove fruitful and will not be allowed to refute the hard core for two reasons: first, they have not "demonstrated, or even shown how to demonstrate, that the explanatory mechanisms [they] postulate are sufficient to account for intelligent behaviour in complex tasks" (Newell & Simon, 1976, p. 120); and second, they have not "been formulated with anything like the specificity of artificial programs" (Newell & Simon, 1976, p. 120).[2] The cognitive-scientific approach to understanding psychological phenomena is best understood by considering its positive heuristics:

explanatory adequacy: experiments on both human behaviour and the neurophysiology of the brain are used to understand the constraints under which mental processes operate and a cognitive theory should account for what is possible within those constraints (Johnson-Laird, 1983; Newell & Simon, 1976).

the doctrine of functionalism: a functional level of description is considered sufficient for the development of theories of cognition; this has two implications: first, so long as the physical substrate provides for an appropriate degree of computational power its physical nature places no constraints on theories of cognition; and second, any scientific theory of cognition may be simulated by a computer program (Chalmers, 1994; Johnson-Laird, 1983; Pylyshyn, 1989).

the criterion of effectiveness: a theory should be defined as an *effective procedure* (*i.e.,* a computer program) to ensure that it takes as little as possible

[1] Most work in artificial intelligence may be classified as applied AI.

[2] Although it is many years since Newell & Simon wrote these words, their thesis remains valid even today.

for granted and any assumptions are clearly stated (Johnson-Laird, 1983; Longuet-Higgins, 1981; Simon & Kaplan, 1989);

empirical evaluation: psychological experiments are required to allow the behaviour of a cognitive model to be evaluated with respect to the human behaviour it purports to explain; as well as goodness of fit to the human data, it is also important to examine discrepancies between the behaviour of the model and the human behaviour as well as any predictions of the model which may not be tested with the current data (Newell & Simon, 1976; Simon & Kaplan, 1989).

The progressive nature of the cognitive-scientific research programme is demonstrated both by its increasing tenacity in modern psychological research and by many specific examples of success such as the accurate prediction of developmental trajectories by cognitive models of language acquisition (see, *e.g.,* Plunkett *et al.*, 1997) and the success of cognitive therapies for anxiety disorders over purely behavioural or biological approaches (see, *e.g.,* Clark & Wells, 1997).

Regarding methodology, Marr (1982) introduced a framework for the understanding of complex information processing systems such as the mind/brain which has proved highly influential in modern cognitive science. Noting that different properties of such systems must be described at different levels of description, Marr isolates three general and relatively autonomous levels at which a description of an information processing system may be placed:

1. the computational theory;
2. the representation and algorithm;
3. the hardware implementation.[3]

The first level deals with the *what* and the *why* of the system. What is the goal of the computation? Why is it appropriate? What is the logic of the strategy by which it can be carried out? At this level, the computational theory attempts to describe the intrinsic nature and computational requirements of a cognitive task through a formal analysis of the various outputs resulting from different inputs. Through understanding the nature of the problem to be solved, appropriate constraints can be placed on the representational and algorithmic levels of the

[3]Pylyshyn (1984) calls these the *semantic* level, the *symbolic* or *syntactic* level and the *biological* or *physical* level respectively. In the interests of clarity the terminology introduced by Marr (1982) is used here.

theory. It is only at the second level of analysis that the question of *how* is addressed; this involves specifying a *representation* for the input and output of the computation and an *algorithm* by which the computation may be achieved. The final level outlined by Marr (1982) concerns the physical realisation of the representation and algorithm. While, on the one hand, the same algorithm may be implemented on a number of different physical substrates, on the other, the choice of hardware may influence the choice of algorithm (between, for example, a serial or parallel algorithm).

One approach to the algorithmic modelling of cognitive processes involves the analysis of a limited and well-circumscribed domain with the goal of finding the exact algorithms underlying the human performance of the task. This has been dubbed the *low road* to understanding cognitive processes (Pylyshyn, 1989). However, for any large-scale problem there is usually a wide range of possible representation schemes and algorithms that may be used. The choices made will depend crucially on the constraints derived from analysing the problem at the computational level (the *high road*). Marr (1982) goes to great lengths to emphasise the importance of the computational theory arguing that the nature of the underlying computations (the second level) depends much more upon the intrinsic computational constraints of the problems to be solved than on the particular hardware mechanisms upon which their solutions are implemented. Speaking of human perception he notes that:

> trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: it just cannot be done.
>
> (Marr, 1982, p. 27)

This three-level analysis of cognitive systems has been criticised by McClamrock (1991) who argues that the transitions between levels conflate two independent types of change. The first describes the level of organisational abstraction of the activity and how functional components of a higher-level explanation may be decomposed into those at a lower level of abstraction. There are clearly many different such levels on which a cognitive system may be described and the actual number of levels of organisation in any particular information processing system "is an entirely empirical matter about that particular system" (McClamrock, 1991, p. 9). The second type of change concerns the types of question asked, or explanations provided, about an information processing system at any particular level of organisation. McClamrock proposes three types of explanation that might be given or questions asked which are roughly analogous to Marr's three levels of description. This interpretation suggests that

there are (at least) two methodological issues to be addressed in any cognitive-scientific research:

1. identify the functional level of description (computational, algorithmic or implementational) of the cognitive system which is to be the prime focus of the research;

2. identify a level of organisational abstraction in the cognitive system which is the prime focus of the research.

It has been argued in §2.2 that the evaluation (by falsification) of scientific theories is crucial to the advance and development of progressive research programmes. In cognitive science, one of the primary purposes of implementing a cognitive theory as a computer program is to allow the detailed and empirical comparison of the behaviour of the program with that of humans on some experimental task (Newell & Simon, 1976; Pylyshyn, 1989). If there exist discrepancies then the model can be improved accordingly and any predictions made by the model can provide suggestions and guidance for further experimental research (Simon & Kaplan, 1989). In the context of modelling music cognition, Desain *et al.* (1998) stress the importance of empirical evaluation:

> proposing a new model ... can hardly be seen as a contribution to the field anymore. Recently a methodology has been emerging in which a working computational model is seen much more as the starting point of analysis and research rather than as the end product ... [it] is thus no longer an aim unto itself but a means to compare and communicate theories between different research communities.
>
> (Desain *et al.*, 1998, p. 153)

2.5 Science and Music

There exist many different motivations for applying AI techniques to the musical domain. These motivations exhibit a wide range of epistemological origins including, for example, those drawn from natural science, engineering, engineering science, the arts and the humanities. This heterogeneity has several sources: first, the fundamental range of motivations existing in AI research (see §2.3); second, the fact that AI techniques are being applied to a domain which is usually studied in the arts and humanities (see §2.2); and third, the

fact that music exists simultaneously as, for example, a physical phenomenon, a psychological phenomenon, an art-form and a performed art.

Given the discussion in §2.2 and §2.3, it will be clear that motivations drawn from different disciplines imply different goals and methodologies for achieving those goals. As a result, the heterogeneity noted above can lead to severe methodological problems in cases where research projects fail to specify the discipline to which they intend to contribute, specify goals appropriate to that discipline and adopt appropriate methodologies for achieving those goals. To illustrate the argument, the application of AI techniques to the generation of music is considered as an example. There exist at least five different motivations that have led to the development of computer programs which compose music and, correspondingly, five distinct activities each with their own goals and appropriate methodologies. The first activity is only tangentially related to music and may be classified as basic AI (see §2.3) since it involves the use of music as an interesting domain for the evaluation of general-purpose AI techniques (see, *e.g.,* Begleiter *et al.*, 2004; Ghahramani & Jordan, 1997). The other activities are discussed in turn.

In the second activity, *algorithmic composition*, computer programs are used to generate novel musical structures, compositional techniques and even genres of music. An example of this motivation is provided by Cope (1991) who developed a system called EMI for algorithmic composition. The motivations and goals are fundamentally artistic since AI techniques are employed as an integral part of the compositional process. As a consequence, there are no methodological constraints placed on the construction of the computer program. Furthermore, there is no need to define any rigorous criteria for success nor to use such criteria in evaluating the program and the compositions. The motivation in other projects is to use AI techniques in *the design of compositional tools* for use by composers. An example of such projects is provided by the research at IRCAM in Paris described by Assayag *et al.* (1999) in which researchers often work together with composers on their products in the task analysis and testing phases of development. Such projects may be classified as applied AI (see §2.3) and should therefore adopt appropriate methodologies from the disciplines of software engineering in the analysis of the task, the design and implementation of the tool and the evaluation of whether the tool satisfies the design requirements.

Other motivations for applying AI techniques to the generation of music are theoretical rather than practical. In the *computational modelling of musical styles*, the goal is to propose and verify hypotheses about the stylistic attributes

defining a corpus of musical works (Ames, 1992; Roads, 1985b). Since the objects of study are existing musical works, this discipline may be considered to be a branch of musicology. The implementation of stylistic hypotheses as a computer program (which can generate music) has two potential advantages (Camilleri, 1992; Sundberg & Lindblom, 1976, 1991). First, while musicology has traditionally adopted speculative methodologies (see §2.2), the computational approach requires that all assumptions included in the theory (self-evident or otherwise) are explicitly and formally stated. The second potential advantage is that the implemented model may be evaluated, and refuted or corroborated, through empirical comparison of the compositions it generates with the human-composed pieces which the theory is intended to describe (see Meredith, 1996). Independent evidence for discriminating between two unrefuted computational theories of a musical style can be obtained by considering the predictions they make about issues commonly addressed in musicology. Examples of such issues include the ability of the models "to distinguish ... structures typical of particular epochs and also ... structures belonging to particular repertoires" (Baroni *et al.*, 1992, p. 187).

The motivations of authors such as Steedman (1984) and Johnson-Laird (1991), discussed in §3.3, were drawn from cognitive science rather than musicology. The distinction is important since "cognitive models need not reflect current music-theoretic constructs, nor must models of musical knowledge have cognitive pretensions" (Desain *et al.*, 1998, p. 152) and the two disciplines differ greatly both in the nature of their goals and the methodologies used to achieve those goals. Following the discussion of cognitive-scientific methodologies in §2.4, there are several advantages to implementing theories of music cognition as computer programs. However, in order to benefit from these advantages, certain methodological practices must be followed. First, a cognitive-scientific model should be based on specific hypotheses, derived from empirical psychological results, which specify the degree of functional organisation they address and kinds of question they pose (see, *e.g.,* Johnson-Laird, 1991). Second, the hypotheses should be evaluated through systematic and empirical attempts to refute them based on comparisons of the behaviour of the implemented model and the human behaviour for which it is intended to account. Once the theory has been corroborated at one level of functional organisation, hypotheses may be formulated and evaluated at a finer level of organisation.

More generally, Cross (1998b) has considered the relevance and utility of different scientific approaches for our understanding of musical phenomena. At one extreme lies the *physicalist* position which holds that the sounds and

structures that we employ and experience in music are wholly determined by the physical nature of sound. Cross rejects the physicalist position because our current understanding of the perception of music indicates that there is not a one-to-one correspondence between physical characteristics of acoustic phenomena (*e.g.,* the frequency and duration of tones) and our perception of those objects.

At the other extreme, Cross (1998b) reviews the deconstructionist or *immanentist* conception of music which is pervasive in current musicological research and which denies the possibility of *any* scientific understanding of music. Cross, however, argues that this is founded on a misconception of scientific methodology as positivist (see §2.2), of scientific knowledge as general (culture independent) and the objects of scientific research being exclusively material. By contrast, a conception of science based on falsificationism (see §2.2) can dispose of many of the objections of the immanentists. In particular, the sophisticated methodological falsificationism of Lakatos (1970) suggests that sufficient weight of change in the background knowledge may contribute to the succession of or radical change in a research programme. Since these research programmes consist partly of local background knowledge and heuristics for change, they are not unsuitable for explaining culturally defined phenomena. Furthermore, the requirement that the scientific evidence be observable does not preclude the scientific study of intentional phenomena, and the provisional and dynamic nature of falsificationism, is consistent with the idea that there are no genuine absolutes.

Having proposed that the arguments of the immanentist position can be overcome, Cross advocates a cognitive-scientific research programme for understanding music. This programme involves the study of all aspects of the musical mind and behaviour at many levels of explanation through theoretical inquiry, formal modelling and empirical experiment. Countless authors have stressed the importance, indeed the necessity, of an interdisciplinary approach to both theoretical and practical research in music. Desain *et al.* (1998), for example, note that the processing and representation of musical structures can provide a common ground for research between disciplines. However, they are careful to distinguish the roles of different disciplines:

> Such structures can be stated formally or informally within music theory, their processing can be investigated by experimental psychology, both of these aspects can be modelled in computer programs and can be given an architectural basis by neuroscience.
>
> (Desain *et al.*, 1998, p. 153)

Each of these disciplines should *embrace* rather than become one with the others (Gjerdingen, 1999a). Research in any discipline may have implications for, or be inspired by, research in any other. However, in any research project it is fundamental to clearly state the motivations involved, the specific goals of the research and the field to which the research contributes in order to allow the adoption of appropriate methodologies for achieving those goals.

2.6 Methodologies for the Present Research

The discussion in §2.4 and §2.5 has provided the foundations of a framework for achieving the aims set out in §1.4. The primary motivations of the current research are cognitive-scientific in character. However, in the development of computational techniques for modelling cognition, subsidiary goals are defined which may be classified as basic AI. In particular, Chapters 6 and 7 present a computational system which is developed and evaluated using methodologies drawn from (basic) AI, rather than cognitive science. In later chapters, this system is applied to the cognitive modelling of music perception and composition. In the present research, the term *cognitive theory* is used to describe an information processing theory of (an aspect of) cognition and the terms *cognitive model* or *computational model* to describe an implemented theory. The term *computational theory* is used to describe cognitive theories which are pitched at the computational (as opposed to the algorithmic or hardware) level(s) of description.

Current understanding of music cognition (including both perception and composition) is currently far less advanced than that of other areas of human psychology (such as visual perception and memory) and detailed algorithmic theories seem a long way off. Since music cognition draws on knowledge and processing in many different domains and at many levels of description, it seems unrealistic to aim towards a purely algorithmic model. Before such an approach becomes possible it will be necessary to understand in more detail the computational level theory describing the overall functional character of the processes involved. As a consequence of these considerations, this research is concerned with computational level theories. Following the discussion in §2.4, the models developed here should be based on specific hypotheses which are stated at a computational level of description, derived from empirical psychological findings concerning music perception and composition, and which identify the level of functional organisation addressed. Any implementational details outwith the defined level of organisational abstraction are taken not as

hypotheses about music cognition but as assumptions necessary for implementing a working model. Any claims made about the computational level theory will concern features of the processing at a level that is abstracted away from the precise algorithmic details.

It has been argued that it is the potential for refutation that distinguishes scientific statements from non-scientific statements. Therefore, any claims made about music cognition must be accompanied by experiments which are capable of refuting those claims. In cognitive science, the implementation of a theory allows the objective evaluation of the behaviour of a model by comparison with the human behaviour it is intended to account for. It also allows predictions to be made about human behaviour based on the behaviour of the model. Therefore, the experimental hypotheses developed in the present research should be evaluated through systematic and empirical attempts to refute them based on comparisons of the behaviour of the implemented models with the human behaviour for which they are intended to account. Part of the contribution made by the present research is the development of a methodology for evaluating hypotheses about music cognition within a computational framework (see Chapter 9). The fields of AI and cognitive science are themselves young disciplines and their application to the musical domain is an even less developed area of investigation: research programmes in music cognition are still in their infant years. The evaluation by falsification of theories in the Lakatosian protective belt of these programmes is crucial so as to build up a theoretical hard core as these theories continue to go unrefuted. Only in this manner can the field begin to build predictive and progressive research programmes.

There are two general approaches to the implementation of cognitive theories of musical competence:

> The first is the knowledge engineering approach, where rules and knowledge are explicitly coded in some logic or grammar ... The second is the empirical induction [or machine learning] approach, where a theory is developed through an analysis of existing compositions.
>
> (Conklin & Witten, 1995, pp. 51–52)

A number of issues arise from the practical difficulties involved in knowledge engineering (Toiviainen, 2000). First, the knowledge and processing involved in many aspects of music cognition are simply not available to conscious introspection. Second, for any reasonably complex domain, it will be practically impossible to capture all the exceptions to any logical system of music description (Conklin & Witten, 1995). An underspecified rule base will not only fail to

describe the genre adequately but will also suffer from bias introduced by the selection of rules by the knowledge engineer:

> the *ad hoc* nature of rule revision is disconcerting: how can the researcher have any confidence that the revisions are the best to propose in the circumstances?
>
> (Marsden, 2000, p. 18)

As discussed in §1.2, the use of expert music-theoretic knowledge in the development of cognitive theories of music perception has been criticised on precisely these grounds.

In the case of a machine learning approach, it is possible to precisely specify the source of the knowledge acquired by the model and the corpus of music over which it may account for observed musical and cognitive phenomena. Since the model acquires its knowledge through exposure to existing music, this approach also offers the possibility of a much more parsimonious account of the influences of (culturally situated) experience on music cognition (see §1.2). It is also important to note that any complete cognitive model of cognitive processing in music perception and composition will also describe how these cognitive skills are acquired and developed (Bharucha & Todd, 1989; Marsden, 2000). The knowledge engineering approach fails to address these issues and often results in inflexible systems which are unable to generalise their knowledge to novel situations. For these reasons, a machine learning approach to the modelling of music and music cognition is adopted in the current research.

2.7 Summary

Methodological and epistemological issues relevant to the present research have been discussed in this chapter. The epistemological nature of scientific knowledge and the distinction between empirical and speculative disciplines was addressed in §2.2 while in §2.3 three branches of AI were introduced along with their characteristic motivations and methodologies. This research falls into the cognitive-scientific tradition of AI research and in §2.4, the dominant methodologies in cognitive science were reviewed. Section 2.5 contained a discussion of methodological concerns which arise specifically in relation to the study of music from the perspective of science and AI. Finally, in §2.6 appropriate methodologies were defined for achieving the objectives of the current research (see §1.4) based on the issues raised in the foregoing sections.

CHAPTER 3

BACKGROUND AND RELATED WORK

3.1 Overview

This chapter contains the background to the modelling techniques used in the present research as well as reviews of previous research which has applied them and related techniques to modelling music and music cognition. In general, the goal of building a computational model of a musical corpus is to develop a grammar which accepts and is capable of assigning structural descriptions to any sequence of symbols in the language used to represent the corpus. However, it is important to be careful when selecting computational methods for both representation and inference since these decisions involve making assumptions about the language in which the corpus is expressed and can impose methodological constraints on the research strategy. In §3.2, context free, finite state and finite context grammars are introduced in terms of the Chomsky containment hierarchy (Hopcroft & Ullman, 1979) and discussed in terms of the languages they can generate, their assumptions and the methodological constraints they impose. In §3.3, previous research on the application of context free (and higher) grammars to the representation and modelling of music is discussed. The application of finite context (or n-gram) grammars and neural networks to modelling music is reviewed in §3.4 and §3.5 respectively. Finally, in §3.6, the application of statistical modelling techniques to various experimentally observed phenomena in music perception is discussed.

3.2 Classes of Formal Grammar

A formal grammar G is a structural description of a formal language consisting of a set of sequences (or strings) composed from some alphabet of symbols (Hopcroft & Ullman, 1979). A grammar itself consists of a tuple (V, T, S, P) where:

- V is a finite set of symbols whose elements are called *non-terminal symbols* or variables;
- T is a finite set of symbols, disjoint from V, whose elements are called *terminal symbols* or constants;
- $S \in T$ is a distinguished symbol called the *initial symbol*;
- P is a set of rewrite rules, or productions, which represent legal transformations of one sequence of terminal and non-terminal symbols into another such sequence.

The language $L(G)$ generated by a grammar G is the subset of T^* which may be rewritten from S in zero or more steps using the productions in P. A sequence is *accepted* by a grammar G if it is a member of $L(G)$.

Noam Chomsky introduced a containment hierarchy of four classes of formal grammar in terms of increasing restrictions placed on the form of valid rewrite rules (Hopcroft & Ullman, 1979). In the following description, $a \in T^*$ denotes a (possibly empty) sequence of terminal symbols, $A, B \in V$ denote non-terminal symbols, $\alpha \in (V \cup T)^+$ denotes a non-empty sequence of terminal and non-terminal symbols and $\beta, \beta' \in (V \cup T)^*$ denote (possibly empty) sequences of terminal and non-terminal symbols.[1]

Type 0 (Unrestricted): grammars in this class place no restrictions on their rewrite rules:

$$\alpha \rightarrow \beta$$

and generate all languages which can be recognised by a universal Turing machine (the recursively enumerable languages).

Type 1 (Context Sensitive): grammars in this class are restricted only in that there must be at least one non-terminal symbol on the left hand side of

[1] See Appendix A for a summary of the notational conventions used in this dissertation.

the rewrite rule and the right hand side must contain at least as many symbols as the left hand side:

$$\beta A \beta' \rightarrow \beta \alpha \beta'$$

and generate all languages which can be recognised by a linear bounded automaton.

Type 2 (Context Free): grammars in this class further restrict the left hand side of their rewrite rules to a single non-terminal symbol:

$$A \rightarrow \alpha$$

and generate all languages which can be recognised by a non-deterministic pushdown automaton.

Type 3 (Finite State or Regular): grammars in this class restrict their rewrite rules further still by allowing only a single terminal symbol, optionally accompanied by a single non-terminal, on the right hand side of their productions:

$$\begin{aligned} A &\rightarrow aB \text{ (right linear grammar) or} \\ &\rightarrow Ba \text{ (left linear grammar)} \\ A &\rightarrow a \end{aligned}$$

and generate all languages which can be recognised by a finite state automaton.

The languages generated by each class of grammar form proper subsets of the languages generated by classes of grammar higher up in the hierarchy. However, as we move up the hierarchy, the complexity of recognition and parsing increases in tandem with the increased expressive power of each class of grammar. In particular, while context free grammars (and those higher in the hierarchy) are capable of capturing phenomena, such as embedded structure, which cannot be captured by finite state grammars, they also bring with them many problems of intractability and undecidability, especially in the context of grammar induction (Hopcroft & Ullman, 1979). It is common, therefore, to try to use a grammar which is only as expressive as the language seems to require.

Context free grammars are typically constructed by hand on the basis of expert knowledge of the language. While methods exist which are, in theory, capable of unsupervised offline learning of probabilistic context free grammars

from unannotated corpora, in practice these methods are subject to limitations which make this task extremely difficult (Manning & Schütze, 1999). Finite state grammars are attractive because they have linear complexity and, unlike context free grammars, are unambiguous with respect to derivation, since each intermediate production has precisely one non-terminal symbol. A Finite State Automaton (FSA) is a quintuple (T, Q, q_0, Q_f, δ) where:

- T is a finite set of input symbols;
- Q is a finite set of states;
- $q_0 \in Q$ is the initial state;
- $Q_f \subset Q$ is the set of final states (or accepting states);
- δ is the state transition function: $\delta : Q \times T \rightarrow Q$.

The FSA is Markovian if, for any state, there is at most one transition for each input symbol. In a probabilistic FSA, δ is defined by a probability distribution over T such that the probabilities of transitions from any given state sum to unity. A Markovian probabilistic FSA is also called a Markov chain.

There is an interesting sub-class of grammar contained within the class of finite state grammars which are known as *finite context* grammars (Bell *et al.*, 1990; Bunton, 1996). Finite context grammars have productions which are restricted to follow the form (Conklin, 1990, p. 40):

$$t^j_{(j-n)+1} \in T^* \quad \rightarrow \quad t^j_{(j-n)+2}$$

In finite context automata, the next state is completely determined by testing a finite portion of length $n-1$ of the end of the already processed portion of the input sequence (Bunton, 1996).[2] The sequence $t^j_{(j-n)+1}$ is called an n-gram which consists, conceptually, of an initial subsequence, $t^{j-1}_{(j-n)+1}$, of length $n-1$ known as the *context* and a single symbol extension, t_j, called the *prediction*. The quantity $n-1$ is the *order* of the n-gram rewrite rule.

A finite context model, or n-gram model, is a database of n-grams with associated frequency counts. The order of an n-gram model is equal to the maximum order n-gram in the database. While there are no existing algorithms for the online induction of finite state models (which are not also finite context

[2] We restrict this discussion to a particular subclass of finite context grammar in which the test is made for membership of a singleton set of sequences and the rewrite rule represents a single symbol extension of the context sequence (Bunton, 1996).

models) while processing an input sequence, it is relatively straightforward to incrementally construct finite context models online by adding new n-grams to the database or incrementing the frequency counts of existing n-grams as new symbols are encountered (Bell *et al.*, 1990). Given a sequence of input symbols, the frequency counts associated with n-grams can be used to return a distribution over T conditioning the estimated probability of a given symbol in the sequence in the context of the $n-1$ preceding symbols. See Chapter 6 for a more detailed discussion of inference using n-gram models.

3.3 Grammars as Representations of Musical Structure

The notion of a musical grammar forms one of the central ideas in music research: "The idea that there is a grammar of music is probably as old as the idea of a grammar itself" (Steedman, 1996, p. 1). Many different types of formal grammar have been applied to many different kinds of problem (including algorithmic composition, musicological analysis and cognitive modelling) in a wide range of musical genres (see Roads, 1985a; Sundberg & Lindblom, 1991, for reviews). Many of these attempts to use grammars to characterise musical styles have used some form of context free grammar. This is, in part, because hierarchical phrase structure is held to be an important feature of (Western tonal) music and, more importantly, the way we perceive music (*e.g.,* Deutsch & Feroe, 1981; Lerdahl, 1988b; Lerdahl & Jackendoff, 1983; Palmer & Krumhansl, 1990).

An example of a computational model following this approach is presented by Johnson-Laird (1991) who used grammatical formalisms to investigate computational constraints on the modelling of improvisational competence in jazz music. Using a corpus of jazz improvisations as examples, Johnson-Laird draws various conclusions concerning what has to be computed to produce acceptable rhythmic structure, chord progressions and melodies in the jazz idiom. These conclusions are stated in terms of constraints imposed on the underlying algorithms generating an improvisation. For example, the analysis suggests that while a finite state grammar (or equivalent procedure) can adequately compute the melodic contour, onset and duration of the next note in a set of Charlie Parker improvisations, its pitch is determined by harmonic constraints derived from a context free grammar modelling harmonic progressions.

Steedman (1984, 1996) describes the use of a context free grammar to describe chord progressions in jazz twelve-bar blues. His goal is to account for the observation that many different chord sequences are perceived by musicians to

be instances of the twelve bar form. In order to achieve this, Steedman developed a categorical grammar (a type of context free grammar) from a theory of tonal harmony due to Longuet-Higgins (1962a,b). Steedman (1996) suggests that this representation "bears a strong resemblance to a 'mental model' in the sense of Johnson-Laird (1983) ... [in that] it builds directly into the representation some of the properties of the system that it represents" (Steedman, 1996, pp. 310–311). He concludes that this computational theory, although less ambitious than that of Johnson-Laird (1991), allows a more elegant description of improvisational competence since it does not rely on substitution into a previously prepared skeleton. However, in using the grammar to generate structural descriptions of blues chord progressions, Steedman was forced to introduce implicit meta-level conventions not explicit in the production rules of the grammar (Wiggins, 1998).

The *Generative Theory of Tonal Music* (GTTM) of Lerdahl & Jackendoff (1983) also warrants discussion since it is probably the best known effort to develop a comprehensive method for the structural description of tonal music. Although it is not a grammar *per se*, it is heavily inspired in many respects by Chomskian grammars. It is, for example, founded on the assumption that a piece of music can be partitioned into hierarchically organised segments which may be derived through the recursive application of the same rules at different levels of the hierarchy. Specifically, the theory is intended to yield a hierarchical, structural description of any piece of Western tonal music which corresponds to the final cognitive state of an experienced listener to that composition.

According to GTTM, a listener unconsciously infers four types of hierarchical structure in a musical surface: first, *grouping structure* which corresponds to the segmentation of the musical surface into units (*e.g.,* motives, phrases and sections); second, *metrical structure* which corresponds to the pattern of periodically recurring strong and weak beats; fourth, *time-span reduction* which represents the relative structural importance of pitch events within contextually established rhythmic units; and finally, *prolongational reduction* reflecting patterns of tension and relaxation amongst pitch events at various levels of structure. According to the theory, grouping and metrical structure are largely derived directly from the musical surface and these structures are used in generating a time-span reduction which is, in turn, used in generating a prolongational reduction. Each of the four domains of organisation is subject to *well-formedness rules* that specify which hierarchical structures are permissible and which themselves may be modified in limited ways by *transformational rules*. While these rules are abstract in that they define only formal possibilities, *pref-*

erence rules select which well-formed or transformed structures actually apply to particular aspects of the musical surface. Time-span and prolongational reduction additionally depend on tonal-harmonic *stability conditions* which are internal schemata induced from previously heard musical surfaces.

When individual preference rules reinforce one another, the analysis is stable and the passage is regarded as stereotypical whilst conflicting preference rules lead to an unstable analysis causing the passage to be perceived as ambiguous and vague. In this way, according to GTTM, the listener unconsciously attempts to arrive at the most stable overall structural description of the musical surface. Experimental studies of human listeners have found support for some of the preliminary components of the theory including the grouping structure (Deliège, 1987) and the metrical structure (Palmer & Krumhansl, 1990).

Roads (1985a) discusses several problems with the use of context free grammars for implementing computational theories of music. In particular, he argues that is not clear that the strict hierarchy characteristic of context free grammars is reconcilable with the ambiguity inherent in music. Faced with the need to consider multiple attributes occurring in multiple overlapping contexts at multiple hierarchical levels, even adding ambiguity to a grammar is unlikely to yield a satisfactory representation of musical context. The use of context sensitive grammars can address these problems to some extent but, as discussed in §3.2, these also bring considerable additional difficulties in terms of efficiency and complexity.

There are, however, several methods of adding some degree of context sensitivity to context free grammars without adding to the complexity of the rewrite rules. An example is the Augmented Transition Network (ATN) which extends a recursive transition network (formally equivalent to a context free grammar) by associating state transition arcs (rewrite rules) with procedures which perform the necessary contextual tests. Cope (1992a,b) describes the use of ATNs to rearrange harmonic, melodic and rhythmic structures in EMI, a simulation of musical thinking in composition. Another example is provided by the pattern grammars developed by Kippen & Bel (1992) for modelling improvisation in North Indian tabla drumming.

However, Roads (1985a) notes, more generally, that musical structure does not yield readily to sharply defined syntactic categories and unique structural descriptions of context free grammars, concluding that:

> In nearly every study [of the application of grammars to musical tasks] the rewrite rule by itself has been shown to be insufficient as a representation for music.

(Roads, 1985a, p. 429)

In conclusion, it is not clear that the power of a context free or context sensitive grammar necessarily brings significant advantages in representing and modelling music. A further problem with the use of such grammars for the computational modelling of music is methodological. Although the induction of context free grammars from unannotated corpora is an ongoing topic of research (Manning & Schütze, 1999), the significant challenges posed by this task mean that, in practice, these grammars are generally hand constructed (see §3.2) and, therefore, require a knowledge engineering approach to modelling music. Several reasons were discussed in §2.6 for strongly preferring an machine learning approach to modelling music and music cognition over the knowledge engineering approach.

3.4 Finite Context Models of Music

N-gram models have been used for music related tasks since the 1950s when they were investigated as tools for composition and analysis (see, *e.g.,* Brooks Jr. *et al.*, 1957; Hiller & Isaacson, 1959; Pinkerton, 1956). Since extensive reviews of this early research exist (Ames, 1987, 1989; Hiller, 1970), we shall focus here on more recent developments in which n-gram models have been applied to a number of musical research tasks including both the development of practical applications and theoretical research. In the former category, we may cite models for computer-assisted composition (Ames, 1989; Assayag *et al.*, 1999; Hall & Smith, 1996), machine improvisation with human performers (Lartillot *et al.*, 2001; Rowe, 1992) and music information retrieval (Pickens *et al.*, 2003) and in the latter, stylistic analysis of music (Dubnov *et al.*, 1998; Ponsford *et al.*, 1999) and cognitive modelling of music perception (Eerola, 2004b; Ferrand *et al.*, 2002; Krumhansl *et al.*, 1999, 2000). In this section, the most relevant aspects of this research to the current work are reviewed with an emphasis on modelling techniques rather than application domain. Many of the interesting aspects of this work reflect attempts to address the limited expressive capacity of n-gram models (see §3.2).

Ponsford *et al.* (1999), for example, have applied simple trigrams and tetragrams to the modelling of harmonic movement in a corpus of 84 seventeenth century *sarabandes*. The objective was to examine the adequacy of simple n-gram models for the description and generation of harmonic movement in the style. The sarabandes were represented in a CHARM-compliant representation scheme (see §5.2.2) and higher order structure was represented in the corpus

through the annotation of events delimiting bars, phrases and entire pieces. A number of pieces were generated within predefined templates (with annotated start, end, bars and phrases) using sequential random sampling of chords from the models with an initial context of length n. Ponsford *et al.* (1999) concluded with an informal stylistic analysis, according to which the generated harmonies were "characteristic of the training corpus in terms of harmony transitions, the way in which pieces, phrases and bars begin and end, modulation between keys and the relation between harmony change and metre" (Ponsford *et al.*, 1999, p. 169). The generation of features such as enharmony, which was not present in the corpus, and weak final cadences was attributed mainly to the use of low order models.

The research most closely related to the present work is that of Darrell Conklin (Conklin, 1990; Conklin & Witten, 1995) who used complex statistical models (see Chapters 6 and 7) to model the soprano lines of 100 of the chorales harmonised by J. S. Bach. A number of strategies were employed to address the problems of imposing a fixed, low order-bound noted by Ponsford *et al.* (1999). In particular, Conklin & Witten combined the predictions of all models with an order less than a fixed threshold in order to arrive at a final prediction (see §6.2.3). Conklin & Witten also combined the predictions of models derived from the entire corpus with transitory models constructed dynamically for each individual composition (see §6.2.4). One of the central features of this work was the development of a framework to extend the application of n-gram modelling techniques to multiple attributes, or *viewpoints*, of a melodic sequence (see §5.2.3). Amongst other things, this framework allows an n-gram model to take advantage of arbitrary attributes derived from the basic representation language, disjunctions of conjunctions of such attributes and contexts representing non-adjacent attributes in a sequence (*e.g.,* using bars, phrases and pieces to denote structurally salient segmentation points as did Ponsford *et al.*, 1999).

Conklin & Witten (1995) developed a number of *multiple viewpoint systems* of the chorale melodies and evaluated them using a number of methods. First, split-sample validation (see §6.3.2) with a training set of 95 compositions and a test set of five compositions was used to compare the performance of the different systems in predicting existing unseen melodies. The performance metric was the cross entropy (see §6.2.2) of the test set given the model. The second means of evaluation was a *generate-and-test* approach similar to that used by Ponsford *et al.* (1999) from which Conklin & Witten concluded that the generated compositions seemed to be "reasonable". Finally, Witten *et al.* (1994) con-

ducted an empirical study of the sequential chromatic pitch predictions made by human listeners on the same test set of compositions (see §8.7.1). The entropy profiles derived from the experimental results for each composition were strikingly similar in form to those generated by the model developed by Conklin & Witten (1995) – the events about which the model was uncertain also proved difficult for humans to predict.

Hall & Smith (1996) have extended the approach used by Conklin & Witten (1995) to a corpus of 58 twelve-bar blues compositions. The aim was to develop a compositional tool that would automatically generate a blues melody when supplied with a twelve-bar blues harmonic structure. In order to model pitch, monogram, digram and trigram models were derived from 48 compositions in the corpus.[3] Separate digram and trigram models were derived for each individual chord occurring in the corpus. Rhythm was represented using an alphabet of short rhythmic patterns (*e.g.,* two semiquavers followed by a quaver) and monogram, digram and trigram models were derived from the training set over this alphabet. When generating rhythms, each generated pattern was screened by a set of symbolic constraints for stylistic suitability. The model was evaluated by asking 198 human subjects to judge which of a pair of compositions (of which one was from the corpus and the other generated by the program) was computer generated (see also §9.2.3). The data consisted of the ten remaining compositions in the corpus and ten compositions randomly selected from the model's output all of which were played to the subjects over a standard harmonic background. Statistical analysis of the results demonstrated that the subjects were unable to distinguish reliably between the human and computer generated compositions.

Reis (1999) has extended the work of Conklin & Witten (1995) in a different direction through the incorporation of psychological constraints in n-gram models. In particular, Reis argues that storing all n-grams (up to a global order bound) occurring in the data is highly inefficient and unlikely to accurately depict the manner in which humans represent melodies. Reis describes a model which segments the data according to perceptual cues such as contour changes or unusually large pitch or duration intervals. The order of the n-grams stored by the model is then determined by the number of events occurring since the previous segmentation point. In the case of ambiguity (*e.g.,* the various segmentation cues do not converge to a single point), all suggested segmentation possibilities are stored. If a novel n-gram is encountered during prediction, the distribution delivered by the variable order model is smoothed with a uni-

[3]Monogram, digram, trigram and tetragram models refer to zeroth, first, second and third order n-gram models respectively.

form distribution over the alphabet. The model also incorporates perceptually guided predictions for events beyond the immediately succeeding one on the basis that humans are often able to anticipate musical events more than one step ahead (*e.g.,* in the case of repeating motifs).

The performance of the model was evaluated on the chorale dataset used by Conklin & Witten (1995) and German folk melodies from the Essen Folk Song Collection (see Chapter 4) using entropy as the performance metric with a split sample experimental design. The results demonstrated that the model failed to outperform that of Conklin & Witten (1995). In spite of this, Reis's work is useful since it addresses the question of which segmentation and modelling strategies work best when model size is limited. In particular, Reis reports the results of an investigation of the predictions of the model when the (perceptually guided) contexts were shifted. On a set of 205 German folk songs he found that while shifts of between one and ten notes always reduced performance relative to non-shifted contexts, shifts of one note and shifts greater than six produced better prediction than other shifts. Reis suggests that the relatively good performance using single note shifts may be explained by a degree of uncertainty as to exactly where a grouping boundary occurs (*i.e.,* is a large melodic interval included in the preceding group or at the beginning of the following group). The improved performance with longer shifts was attributed to the fact that the average length of the suggested segments was 6.7 notes. Regarding multiple-step-ahead prediction, Reis found that predictions made further in advance lead to lower prediction performance and suggested that this was due to the lack of matching contexts for these higher order predictions.

Cypher (Rowe, 1992) is an interactive music system designed as a compositional tool whose compositional module uses an n-gram model. While Cypher does not have cognitive aspirations, it does use perceptual cues, such as discontinuities in pitch, duration, dynamics and tempo, to determine phrasing boundaries. Events occurring on metric pulses are also given more weighting as segmentation points. These various cues are combined using a predefined weighting to give a value for each event. If this value exceeds a certain threshold, the event is classed as a segmentation boundary. The order of the model is given by the number of events that have occurred since the previous segmentation point. In contrast to the work of Reis (1999), Cypher employs *a priori* tonal knowledge to aid segmentation and a fixed weighting in the combination of perceptual cues to arrive at one ambiguous segmentation.

More distantly related approaches are also relevant here since they have been used to tackle the same basic task – the prediction of an event given

a context of immediately preceding events. Assayag, Dubnov and their colleagues (Assayag *et al.*, 1999; Dubnov *et al.*, 1998; Lartillot *et al.*, 2001) have experimented with using an incremental parsing (IP) algorithm based on the Ziv-Lempel dictionary compression algorithm (Ziv & Lempel, 1978) in the modelling of musical styles. The incremental parsing algorithm adaptively builds a dictionary of sequences as follows. For each new event, it appends the event to the current contender for addition to the dictionary (initially the empty sequence). If the resulting sequence occurs in the dictionary, the count associated with that dictionary entry is incremented; otherwise the sequence is added to the dictionary and the current contender is reset to the empty sequence. The algorithm then progresses to the next input event. During prediction, an order bound is specified and Maximum Likelihood estimates (see §6.2.1) are used to predict events in the current context. When the context does not appear in the dictionary, the longest suffix of that context is tried. The IP algorithm has been used successfully, with certain improvements, for the classification of polyphonic music by stylistic genre (Dubnov *et al.*, 1998) and for polyphonic improvisation and composition (Assayag *et al.*, 1999; Lartillot *et al.*, 2001).

Lartillot *et al.* (2001) have also experimented with another technique for constructing variable order n-gram models called *Prediction Suffix Trees* (PST). The algorithm for constructing a PST, originally developed by Ron *et al.* (1996) and used by Lartillot *et al.* (2001), is offline and operates in two stages. In the first stage, a suffix tree is constructed from all subsequences of the input sequence less than a global order bound. In the second stage, each node in the tree is examined and pruned unless for some symbol in the alphabet, the estimated probability of observing that symbol at the node exceeds a threshold value and is significantly different from the estimated probability of encountering that symbol after the longest suffix of the sequence represented by that state. Lartillot *et al.* (2001) have derived PSTs from music in a range of different styles and generated new pieces with some success.

Triviño-Rodriguez & Morales-Bueno (2001) have developed an extended PST model which can compute next symbol probabilities on the basis of a number of event attributes (*cf.* Conklin & Witten, 1995). These Multiattribute Prediction Suffix Graphs (MPSGs) were used to model the chorale melodies used as data by Conklin & Witten (1995) in terms of chromatic pitch, duration and key signature. New chorale melodies were generated from the model using sequential random sampling. A Kolmogorov-Smirnov test failed to distinguish the monogram and digram distributions of pitches in the generated pieces from those in the training corpus. Furthermore, Triviño-Rodriguez & Morales-Bueno

(2001) performed a listening test, with 52 subjects, each of whom was asked to listen to one generated melody and one melody from the training set and classify them according to whether or not they were generated by the model (see also §9.2.3). The results showed that the listeners were able to correctly classify melodies in just 55% of cases.

3.5 Neural Network Models of Music

Mozer (1994) argues that the use of *n*-gram models (such as those described in §3.4) suffer from two fundamental problems in terms of modelling music: first, non-consecutive events cannot be predicted without knowledge of the intervening notes; and second, the symbolic representation used does not facilitate generalisation from one musical context to perceptually similar contexts. In order to overcome these problems, Mozer developed a model based on a Recurrent Artificial Neural Network (RANN - Elman, 1990) and used psychoacoustic constraints in the representation of pitch and duration. In particular, the networks operated over multidimensional spatial representations of pitch (which emphasised a number of pitch relations including pitch height, pitch chroma and fifth relatedness, Shepard, 1982) and duration (emphasising such relations as relative duration and tuplet class).

In contrast to n-gram models, which acquire knowledge through unsupervised induction, these neural networks are trained within a supervised regime in which the discrepancy between the activation of the output units (the expected next event) and the desired activation (the actual next event) is used to adjust the network weights at each stage of training.[4] When trained and tested on sets of simple artificial pitch sequences with a split-sample experimental paradigm, the RANN model outperformed digram models. In particular, the use of cognitively motivated multidimensional spatial representations led to significant benefits (over a local pitch representation) in the training of the networks. However, the results were less than satisfactory when the model was trained on a set of melodic lines from ten compositions by J. S. Bach and used to generate new melodies: "While the local contours made sense, the pieces were not musically coherent, lacking thematic structure and having minimal phrase structure and thematic organisation" (Mozer, 1994, p. 273). The neural network architecture appeared unable to capture the higher level structure in these longer pieces of music.

[4] Page (1999) criticises RANNs as models of melodic sequence learning for a variety of reasons including their use of supervised learning.

One approach to addressing the apparent inability of RANNs to represent recursive constituent structure in music involves what is called auto-association. An auto-associative network is simply one which is trained to reproduce on its output layer a pattern presented to its input layer, generally forming a compressed representation of the input on its hidden layer. For example, training a network with eight-unit input and output layers separated by a three-unit hidden layer with the eight 1-bit-in-8 patterns typically results in a 3-bit binary code on the hidden units (Rumelhart *et al.*, 1986). Pollack (1990) introduced an extension of auto-association called Recursive Auto-Associative Memory (RAAM) which is capable of learning fixed-width representations for compositional tree structures through repeated compression. The RAAM architecture consists of two separate networks: first, an encoder network which constructs a fixed-dimensional code by recursively processing the nodes of a symbolic tree from the bottom up; and second, a decoder network which recursively decompresses this code into its component parts until it terminates in symbols, thus reconstructing the tree from the top down. The two networks are trained in tandem as a single auto-associator.

Large *et al.* (1995) examined the ability of RAAM to acquire reduced representations of Western children's melodies represented as tree structures according to music-theoretic predictions (Lerdahl & Jackendoff, 1983). It was found that the trained models acquired compressed representations of the melodies in which structurally salient events are represented more efficiently (and reproduced more accurately) than other events. Furthermore, the trained network showed some ability to generalise beyond the training examples to variant and novel melodies although, in general, performance was affected by the depth of the tree structure used to represent the input melodies with greater degrees of hierarchical nesting leading to impaired reproduction of input melodies. However, the certainty with which the trained network reconstructed events correlated well with music-theoretic predictions of structural importance (Lerdahl & Jackendoff, 1983) and cognitive representations of structural importance as assessed by empirical data on the events retained by trained pianists across improvised variations on the melodies.

Other researchers have sought to address the limitations of neural network models of music by combining them with symbolic models (Papadopoulos & Wiggins, 1999). HARMONET (Hild *et al.*, 1992) is an example of such an approach. The objective of this research was to approximate the function mapping chorale melodies onto their harmonisation using a training set of 400 four-part chorales harmonised by J. S. Bach. The problem was approached by decom-

posing it into sub-tasks: first, generating a skeleton structure of the harmony based on local context; second, generating a chord structure consistent with the harmonic skeleton; and finally, adding ornamental quavers to the chord skeleton. Neural networks were used for the first and third sub-tasks and a symbolic constraint satisfaction approach was applied to the second sub-task. Hild *et al.* (1992) found that this hybrid approach allowed the networks to operate within structural constraints, relieving them of the burden of learning structures which may be easily expressed symbolically. The resulting harmonisations were judged by an audience of professional musicians to be on the level of an improvising organist. Furthermore, independent support has been found for the predictions made by the model. For example, Hörnel & Olbrich (1999) used HARMONET to predict that a certain chorale harmonisation (attributed to Bach) had not been composed by Bach himself since it coincided to a greater degree with the analysis of an earlier style of harmonisation than with the analysis of Bach's own chorale harmonisations. This prediction was confirmed in the musicological literature.

In an extension to this research, Hörnel (1997) examined the modelling of melodic variation in chorales harmonised by J. Pachelbel. The learning task was performed in two steps. First, given an input melody, HARMONET was used to invent a chorale harmonisation of the melody. In the second stage, a multi-scale neural network was used to provide melodic variations for one of the voices. The latter task was considered at two different scales: in the first stage, a neural network learns the structure of melodies in terms of sequences of four note *motif classes* while in the second stage, another neural network learns the implementation of abstract motif classes as concrete notes, depending on the generated harmonic context. The abstract motif classes were discovered by classifying melodic motifs in the training set with an unsupervised clustering algorithm. A representation was developed for motifs in which a note is represented by its interval to the first motif note in terms of interval size, direction and octave. In informal stylistic analyses of the melodic variations generated by the system, Hörnel (1997) concluded that the modelling of abstract motif classes significantly aided the network in producing harmonically satisfying and coherent variations.

3.6 Statistical Modelling of Music Perception

The n-gram and neural network models described in §3.4 and §3.5 acquire knowledge about a musical style by inducing statistical regularities in existing

corpora of music in that style. They can use this acquired knowledge to model novel works in the style (either in a synthetic or analytical capacity) in terms of the estimated probabilities of events occurring in a given musical context. Meyer (1956, 1957) argues that humans also induce statistical knowledge of a musical style through listening to music in that style and, furthermore, that they bring this acquired knowledge to bear when listening to novel pieces in the style (see also Chapter 8). Furthermore, there is experimental evidence to support the hypothesis that humans internalise a complex system of regularities about music in a given stylistic tradition.

Krumhansl & Kessler (1982) derived the perceived strength of scale degrees as experimentally quantified *key profiles* using a *probe tone* experimental paradigm. Ten musically trained Western listeners were asked to supply ratings of how well a variable probe tone fitted, in a musical sense, with a antecedent musical context. The probe tones in Krumhansl & Kessler's experiments were all 12 tones of the chromatic scale which were presented in a variety of contexts, including complete diatonic scales, tonic triads and a number of chord cadences in both major and minor keys and using a variety of tonal centres. The results exhibited high consistency across contexts and both inter- and intra-subject consistency were also high. Furthermore, substantially the same patterns were found for the different major and minor contexts once the ratings had been transposed to compensate for the different tonics. Krumhansl & Kessler (1982) derived quantitative profiles for the stability of tones in major and minor keys from the experimental data. Since these profiles substantially conform to music-theoretic accounts of decreasing relative stability from the tonic, through the third and fifth scale degrees, the remaining diatonic scale degrees and finally the non-diatonic tones, they have been cited as evidence for a perceived *tonal hierarchy* of stability (Krumhansl, 1990).

Krumhansl (1990, ch. 3) reports a number of case studies in which the key profiles of Krumhansl & Kessler (1982) are found to exhibit strong correlations with the monogram distributions of tones in a variety of musical styles including the vocal melodies in songs and arias composed by Schubert, Mendelssohn, Schumann, Mozart, J. A. Hasse and R. Strauss (Knopoff & Hutchinson, 1983; Youngblood, 1958) and the melodies of nursery tunes (Pinkerton, 1956). In a further study, Krumhansl (1990) found an even stronger relationship between the key profiles and the monogram distribution of duration/pitch pairs reported in a polyphonic analysis of the first of Schubert's *Moments Musicaux*, op. 94, no. 1, (Hughes, 1977). Finally, Krumhansl (1990) argues that the statistical usage of tones in existing musical traditions is the dominant influence on per-

ceived tonal hierarchies, the influence of factors such as acoustic consonance being rather small, suggesting that tonal hierarchies are primarily acquired through learning.[5]

As discussed in §8.2.3 and §8.3.2, experimental research using fixed low-order n-grams has demonstrated the sensitivity of listeners to statistical regularities in music and the influence of these regularities on segmentation (Ferrand *et al.*, 2002; Saffran *et al.*, 1999), stylistic judgements (Vos & Troost, 1989) and expectation for forthcoming tones in a variety of melodic contexts (Eerola *et al.*, 2002; Krumhansl *et al.*, 1999; Oram & Cuddy, 1995). In a cross-cultural context, Castellano *et al.* (1984) collected the probe tone ratings of Indian and Western listeners in the context of North Indian rāgs and found that the responses of both groups of listeners generally reflected the monogram distribution of tones in the musical contexts.

Research has also addressed the question of whether statistical knowledge is acquired and employed in the perception of harmonic relations. Bharucha (1987), for example, developed a connectionist model of harmony based on a sequential feed-forward neural network similar to those used by Mozer (1994). The model accurately predicts a range of experimental findings including memory confusions for target chords following a context chord (Bharucha, 1987) and facilitation in priming studies (Bharucha & Stoeckig, 1986, 1987). In these latter studies, target chords which are musically related to a prior context chord were found to be processed more quickly and accurately than unrelated target chords. Speed and accuracy of judgements varied monotonically with the distance, around the circle of fifths, of the chord from the context inducing chord even when harmonic spectra shared by the context and target chord were removed. The network model learnt the regularities of typical Western chord progressions through exposure and the representation of chord proximity in the circle of fifths arose as an emergent property of the interaction of the network with its environment.

3.7 Summary

In this chapter, the background to the modelling techniques used in the current research has been presented and previous research which has applied them

[5] Krumhansl (1990) reviews a case study conducted by L. K. Miller (referred to in Miller, 1987) of a developmentally disabled boy who, in spite of being musically untrained, exhibited a striking ability to reproduce previously unheard melodies on the piano. When asked to reproduce short preludes, the renditions he produced, while they deviated considerably from the originals, preserved almost exactly the monogram distribution of tone frequencies in the original prelude.

and related techniques to modelling music and music cognition has been reviewed and discussed. In §3.2, context free, finite state and finite context grammars were introduced in terms of the Chomsky containment hierarchy and discussed in terms of the languages they can generate, their assumptions and the methodological constraints they impose. In §3.3, previous research on the application of context free (and higher) grammars to music was summarised and discussed. In particular, while context free (and higher) grammars can be useful in identifying computational constraints on musical competence, they suffer from the problems of inadequate modelling of musical context and the difficulty of adopting a machine learning approach (advocated in §2.4). The application of finite context (or n-gram) grammars and neural networks to modelling music was reviewed in §3.4 and §3.5 respectively. While these models also suffer from an inability to adequately model musical languages, they have advantages in terms of being relatively straightforward to induce from a corpus of data and several approaches to addressing their inherent limitations have been discussed. Finally, in §3.6, research was reviewed which demonstrates the utility of statistical modelling techniques in accounting for a variety of experimentally observed phenomena in music perception.

CHAPTER 4

MUSIC CORPORA

4.1 Overview

In this chapter, issues concerning the selection of data are discussed and the corpora of music used in the present research are described.

4.2 Issues Involved in Selecting a Corpus

There are several criteria that should be borne in mind when choosing a corpus of data for a machine learning approach to modelling music. First, various pragmatic requirements must be met. The compositions should be easily accessible in some electronic format and should be out of copyright. Furthermore, it should be possible to derive all the required information from the original electronic format and transfer this information into the representation scheme used (see §5.3) in a relatively straightforward manner.

Apart from purely pragmatic factors, several issues arise from the focus of the present research on statistical induction of regularities in the selected corpora. First, a corpus should be large enough to support the induction of statistical and structural regularities in the music from which it is comprised. Second, if the music making up a corpus (regardless of its size) exhibits extreme structural diversity then that corpus will afford few regularities which may be exploited in training the models. The structural coherence of a corpus is also important in evaluating compositions generated by a statistical model trained on that corpus (see Chapter 9). Judging the success of a generated composi-

tion in the context of the stylistic norms of the corpus will be very much harder when the corpus exhibits high degrees of diversity and irregularity amongst the individual compositions from which it is comprised. The risk of choosing corpora which exhibit little structural coherence may be potentially reduced by selecting compositions for inclusion in a given corpus according to a common historical period, geographical region, culturally or socially defined musical tradition, specific composer and so on. Third, in order to ensure the ecological validity of the research, that is to ensure that the results pertain to a "real-world" phenomenon, a corpus should consist of entire compositions drawn from existing musical traditions. Finally, a corpus should be stylistically simple enough to enable modelling the data in a relatively complete manner but should exhibit enough complexity to require more than a trivial modelling approach.

4.3 The Datasets

Several corpora of musical data have been chosen to satisfy the criteria discussed in §4.2. In order to reduce the complexity of the task while working with compositions drawn from existing musical traditions, all the datasets selected contain purely monophonic compositions (see §1.4). Stylistically the datasets chosen consist of folk and hymn music and were all obtained in the `**kern` format (Huron, 1997) from the *Centre for Computer Assisted Research in the Humanities* (CCARH) at Stanford University, California (see `http://www.ccarh.org`) and the *Music Cognition Laboratory* at Ohio State University (see `http://kern.humdrum.net`).

The datasets used in the current research are as follows (see Table 4.1 for a summary). The first is a collection of 152 folk songs and ballads from Nova Scotia, Canada collected between 1928 and 1932 by Helen Creighton (1966). The dataset was encoded in the `**kern` format by Craig Sapp and is freely available from the *Music Cognition Laboratory* at Ohio State University. The second dataset used is a subset of the chorale melodies harmonised by J. S. Bach (Riemenschneider, 1941). A set of 185 chorales (BWV 253 to BWV 438) has been encoded by Steven Rasmussen and is freely available in the `**kern` format from CCARH. The remaining datasets come from the Essen Folk Song Collection (EFSC – Schaffrath, 1992, 1994) which consists of a large number of (mostly) European and Chinese folk melodies collected and encoded under the supervision of Helmut Schaffrath at the University of Essen in Germany between 1982 and 1994. A dataset containing 6251 compositions in the collection encoded in the `**kern` format is published and distributed by CCARH (Schaffrath, 1995)

ID	Description	Melodies	Events	E/M	Pitches
1	Canadian folk ballads	152	8553	56.270	25
2	Chorale melodies	185	9227	49.876	21
3	Alsatian folk songs	91	4496	49.407	32
4	Yugoslavian folk songs	119	2691	22.613	25
5	Swiss folk songs	93	4586	49.312	34
6	Austrian folk songs	104	5306	51.019	35
7	German folk songs (`kinder`)	213	8393	39.403	27
8	Chinese folk songs	237	11056	46.650	41
9	German folk songs (`fink`)	566	33087	58.457	37
Total		1760	87395	49.656	45

Table 4.1: Melodic datasets used in the present research; the columns headed E/M and Pitches respectively indicate the mean number of events per melody and the number of distinct chromatic pitches in the dataset.

while an additional dataset of 2580 Chinese folk melodies is available on request from the *Music Cognition Laboratory* at Ohio State University. The six datasets from the EFSC used in the present research contain respectively 91 Alsatian folk melodies, 119 Yugoslavian folk melodies, 93 Swiss folk melodies, 104 Austrian folk melodies, 213 German folk melodies (dataset `kinder`), 566 German folk melodies (dataset `fink`) and 237 Chinese folk melodies (selected from dataset `shanxi`). See Appendix B for an example of the `**kern` encoding of one of the folk songs from the EFSC.

Each dataset is assigned a positive integer as an identifier as shown in Table 4.1 and will be referred to henceforth by this identifier. Table 4.1 also contains more detailed information about each dataset, including the number of melodies and events contained in the dataset as well as the mean number of events per melody. Since the present research focuses on the pitch structure of the melodies in the nine corpora, Table 4.1 also lists the number of distinct chromatic pitches from which each dataset is composed.

4.4 Summary

In this chapter, issues concerning the selection of data were discussed and the corpora of music used in the present research were described.

CHAPTER 5

THE REPRESENTATION OF MUSICAL STRUCTURE

5.1 Overview

This chapter presents the representation scheme used in the current research. The manner in which knowledge is represented is crucial to the success of an AI system. As an example, consider simple arithmetic where some calculations will be extremely easy using a decimal representation but harder when using the Roman numeral system, while for other calculations the converse will be true (Marr, 1982). The choice of an appropriate representation scheme is dependent on the type of information processing that is to be carried out and "search can be reduced or avoided by selecting an appropriate problem space" (Newell & Simon, 1976, p. 125). Furthermore, since the present research is cognitive-scientific, the representation scheme should also be constrained by current understanding of human cognitive representations of music. Following Harris *et al.* (1991), the *musical surface* (Jackendoff, 1987) is taken to correspond to the lowest level of musical detail which is of interest; in this case, the discrete properties of discrete musical events at the note level. Lower-level acoustic phenomena are not considered in this research. This decision may be justified by noting that many aspects of music theory, perception and composition operate on musical phenomena defined at this level (Balzano, 1986b; Bharucha, 1991; Krumhansl, 1990; Lerdahl, 1988a).

Wiggins *et al.* (1993) introduce two orthogonal dimensions along which representation systems for music may be classified: *expressive completeness* and *structural generality*. The former refers to the range of raw musical data that

can be represented while the latter refers to the range of high-level structures that may be represented and manipulated. For example, waveforms have high expressive completeness but low structural generality while traditional scores have high structural generality but restricted expressive completeness. Different tasks will place different emphasis on each of the properties; archiving, for example, places a stress on accuracy of storage and recall, and requires high expressive completeness, while for analysis and composition, structural generality is more important. Of primary concern in the present research is to choose a representation scheme with high structural generality. In particular, a significant challenge faced in modelling musical phenomena arises from the need to represent and manipulate many different features of the musical surface in tandem.

The chapter is organised as follows. Section 5.2 contains a review of existing frameworks for the symbolic representation of music which form the basis for the representation scheme used in the current research. These include the Generalised Interval Systems of Lewin (1987), CHARM (Harris *et al.*, 1991; Smaill *et al.*, 1993; Wiggins *et al.*, 1989) and the multiple viewpoints framework (Conklin, 1990; Conklin & Witten, 1995). The representation scheme used in the present research draws on ideas from CHARM and, especially, the multiple viewpoints framework both of which draw on different aspects of Lewin's Generalised Interval Systems. The preprocessing of the data (described in Chapter 4) and the basic event representation employed in the current research are presented in §5.3. Finally, in §5.4 the multiple viewpoint framework developed in the current research is described in detail and the individual attribute types implemented are motivated in terms of previous research on music cognition and the computational modelling of music.

5.2 Background

5.2.1 Generalised Interval Systems

Lewin (1987) takes as his goal the formal description of various kinds of musical space and, in particular, the precise characterisation of various distance metrics between points in such spaces. He develops a mathematical model, called a *Generalised Interval System* or *GIS*, to describe our intuitions about the relationships between points in musical spaces and discusses two methods for deriving new GISs from existing ones.

Formally, a GIS is an ordered triple $(S, IVLS, int)$ where S is a set of elements defining the musical space, $IVLS$ is a mathematical group consisting of a

set of intervals between the elements of S and an associative, binary operation $*$ on this set, and $int : S \times S \rightarrow IVLS$ is a function mapping pairs of elements in S onto intervals in $IVLS$, subject to Conditions 5.1 and 5.2.

$$\forall p, q, r \in S,\ int(p, q) * int(q, r) = int(p, r) \tag{5.1}$$

$$\begin{aligned} &\forall p \in S, i \in IVLS, \\ &\quad \exists q_1 \in S : int(p, q_1) = i \ \wedge \\ &\quad \forall q_2 \in S, int(p, q_2) = i \Rightarrow q_1 = q_2 \end{aligned} \tag{5.2}$$

Together with the group structure of $IVLS$, Condition 5.1 ensures the existence of an identity interval $e \in IVLS$ such that $\forall p \in S, int(p, p) = e$ and the existence of an inverse interval for each element in $IVLS$ such that $\forall p, q \in S, int(p, q) = int(q, p)^{-1}$. Condition 5.2, on the other hand, ensures that the space S is large enough to contain all theoretically conceivable elements: if we can conceive of an element $p \in S$ and an interval $i \in IVLS$ then we can conceive of a unique element q which lies the interval i from p. Of the many different GISs discussed by Lewin (1987), we shall consider three examples respectively involving a pitch space, a space of temporal points and a space of event durations.

GIS 2.1.2 In this GIS, S is an equal-tempered chromatic scale extended infinitely up and down, $IVLS$ is the group of integers under addition and the function $int(p, q)$, given any two pitches p and q, returns the number of semitones up from p to q. A negative member of $IVLS$ indicates a downward interval of the specified number of semitones.

GIS 2.2.1 In this GIS, S is a set of regularly spaced time points extending indefinitely both forwards and backwards, $IVLS$ is the group of integers under addition and $int(p, q)$ is the difference in terms of temporal units between p and q.

GIS 2.2.3 In this GIS, S is a set of durations measuring a temporal span in time units, $IVLS$ is some multiplicative group of positive real numbers and $int(p, q)$ is the quotient of the durations p and q, (*i.e.,* $\frac{q}{p}$). For example, if p spans four time units and q spans three time units $int(p, q) = \frac{3}{4}$. The elements of $IVLS$ will depend on the proportions amongst durations that we wish to allow.

Having considered some examples of GISs, we can now appreciate Lewin's argument that a GIS is capable of capturing many of our basic intuitions about musical spaces. Thus to the extent that we intuit such musical spaces, we also intuit intervals in connexion with them. The use of a mathematical group to represent these intervals captures the following intuitions:

- the intervals can be composed or combined to generate other intervals (see Condition 5.1);
- this composition of intervals is associative, $a * (b * c) = (a * b) * c$;
- there exists an identity interval I such that $a * I = I * a = a$;
- each interval has an inverse interval such that $int(p, q) = int^{-1}(q, p)$;
- if we can conceive of a point $p \in S$ and an interval $a \in IVLS$ then the musical space S must contain the point that lies the interval a from p (see Condition 5.2).

These properties are logically implied by the group structure of $IVLS$ and Conditions 5.1 and 5.2.

Lewin (1987) provides formal accounts of two ways in which new GIS structures may be created from existing ones. The first of these involves the construction of a *quotient GIS* formed by defining and applying a congruence to an existing GIS. Let $(S_1, IVLS_1, int_1)$ be a GIS and $CONG$ be any congruence on $IVLS_1$. An equivalence relation $EQUIV$ is induced on S_1 by declaring $s, s' \in S_1$ to be equivalent whenever $int_1(s, s')$ is congruent to the identity I in $IVLS_1$. Let S_2 be the quotient space $S_1 \setminus EQUIV$ and $IVLS_2$ be the quotient group $IVLS_1 \setminus CONG$. The function $int_2 : S_2 \times S_2 \to IVLS_2$ is well defined by the following method: given equivalence classes $p, q \in S_2$, the value $int_2(p, q)$ is that member of $IVLS_2$ to which $int_1(q_1, q_2)$ belong whenever q_1 and q_2 are members of p and q respectively. Furthermore, $(S_2, IVLS_2, int_2)$ is itself a GIS.

To give an example of the construction of a quotient GIS, let $(S_1, IVLS_1, int_1)$ be GIS 2.1.2 described above and $CONG$ be the relation on $IVLS_1$ that makes a congruent to a' whenever the intervals differ by any integral multiple of 12 semitones. Then the quotient GIS $(S_2, IVLS_2, int_2)$ constructed by the above method has these components: S_2 is the set of 12 pitch-classes, $IVLS_2$ is the set of integers under $\mathrm{mod}12$ addition and $int_2(p, q)$ is the reduction modulo 12 of the integer $int_1(p_1, q_1)$ where p_1 and q_1 are any pitches belonging to the pitch classes p and q respectively.

The second means of deriving new GIS structures involves the construction of a *product GIS* from two existing GISs. Given $\mathrm{GIS}_1 = (S_1, IVLS_1, int_1)$

and GIS$_2 = (S_2, IVLS_2, int_2)$, the direct product of GIS$_1$ and GIS$_2$, denoted by GIS$_1 \otimes$ GIS$_2$, is that GIS$_3 = (S_3, IVLS_3, int_3)$ constructed as follows:

- $S_3 = S_1 \times S_2$;
- $IVLS_3 = IVLS_1 \otimes IVLS_2$;
- $int : S_3 \times S_3 \rightarrow IVLS_3$ is given by the rule: $int_3((p_1, p_2), (q_1, q_2)) = (int_1(p_1, q_1), int_2(p_2, q_2))$.

As an example, let GIS$_1$ be GIS 2.2.1 and GIS$_2$ be GIS 2.2.3, as described above. Then GIS$_3$ = GIS$_1 \otimes$ GIS$_2$ consists of:

- $S_3 = S_1 \times S_2$; we can conceive of a point (t, d) in this space modelling an event that begins at time-point t and extends for a duration of d time-points thereafter;
- $IVLS_3 = IVLS_1 \otimes IVLS_2$; each interval consists of a pair (a_1, a_2) where a_1 represents the number of time-units between two time-points and a_2 represents a quotient of two durations;
- $int_3((t, d), (t_1, d_1)) = (int_1(t, t_1), int_2(d, d_1))$; if $int_3((t, d), (t_1, d_1)) = (a, b)$ this tells us that (t_1, d_1) occurs a time units after (t, d) and lasts for b times the duration of (t, d).

In further developments of the approach, Lewin (1987) defines formal generalised analogues of transposition and inversion operations using GIS structures and eventually incorporates GISs into a more general formulation using transformational networks or graphs.

5.2.2 CHARM

CHARM (Common Hierarchical Abstract Representation for Music) is intended to provide a logical specification of an abstract representation of music for use in a wide range of areas including composition, analysis and archiving. An abstract, logical scheme allows the flexible representation of many different kinds of musical structure at appropriate levels of generality for any particular task (Wiggins *et al.*, 1989), independent of the particular style, tonal system, tradition or application under consideration (Smaill *et al.*, 1993).

5.2.2.1 Representing Events

The fundamental level of representation in CHARM is the *event* which in general corresponds to the musical note; the internal structure of individual tones and timbres is not represented (Harris *et al.*, 1991; Wiggins & Smaill, 2000). Attributes such as pitch and duration are not represented in a continuous fashion due to practical considerations (Wiggins *et al.*, 1993). Although CHARM may represent performed, perceived or transcribed musical objects, the original formulation focused on performed musical objects since scores were viewed as performance instructions which may have several different interpretations (Harris *et al.*, 1991; Wiggins *et al.*, 1993). Consequently, features such as time signatures and key signatures were not explicitly represented.

Specification The internal structure of an event is represented by abstract data types for pitch (and pitch interval), time (and duration), amplitude (and relative amplitude) and timbre. Therefore, the abstract event representation is the Cartesian product:

$$\text{Pitch} \times \text{Time} \times \text{Duration} \times \text{Amplitude} \times \text{Timbre}$$

Since all of these attributes (except timbre) have the same internal structure, we shall only consider the representation of Time. The objects of interest are points in time and time-intervals (or durations) and the Time and Duration types are associated with sets containing possible values for these event attributes (Wiggins *et al.*, 1989). The abstract data-type for Duration is associated with:

- a distinguished symbol for the zero duration;
- an operation $add_{dd} : \ Duration \times Duration \rightarrow Duration$ and its inverse sub_{dd};
- an ordering given by the typed equivalents of arithmetic comparison functions (*e.g.*, $\leq$, $\geq$, $=$, $\neq$).

such that these make it a linearly ordered Abelian group. To allow the computation of durations from pairs of times, times from pairs of times (and pairs of durations) and so on, the specification requires the existence of functions add_{xy} and sub_{xy} where x and y are one of $\{t, d\}$:

$$
\begin{aligned}
add_{dd}: &\quad Duration \times Duration \rightarrow Duration \\
add_{td}: &\quad Time \times Duration \rightarrow Time \\
sub_{dd}: &\quad Duration \times Duration \rightarrow Duration \\
sub_{td}: &\quad Time \times Duration \rightarrow Time \\
sub_{tt}: &\quad Time \times Time \rightarrow Duration
\end{aligned}
$$

The function sub_{tt} allows us to compute the time interval (duration) between two time points such that:

- $sub_{tt}(p, q) = 0 \Leftarrow p = q$;
- $sub_{tt}(p, q) + sub_{tt}(q, r) = sub_{tt}(p, r)$;
- $sub_{tt}(p, q) = -sub_{tt}(q, p)$.

Formally, this means that Time is a commutative GIS in the sense of Lewin (1987) with extra properties. According to Harris *et al.* (1991), the abstract data types for the other event attributes (except timbre) have the same properties modulo renaming.

Implementation All events must be associated with a unique identifier and a tuple of event attributes making event tuples of the form:

```
event(Identifier,Pitch,Time,Duration,Amplitude,Timbre)
```

Each data type *X* (pitch, pitch interval, time, duration and so on) must have an associated unary selector function *getX* which returns the appropriate component of the event tuple associated with the identifier provided as an argument. There must also be a unary selector function *putEvent* which returns the identifier associated with any event tuple passed as an argument (Harris *et al.*, 1991). These functions are intended to be used with a database in which the events are stored.

5.2.2.2 Representing Higher-order Structures

Harris *et al.* (1991) note that while representation schemes for music must allow higher-level forms to be introduced hierarchically, they must not impose any one set of groupings on the user and should ideally allow the simultaneous assignment of different higher-level structures to any set of musical events.

This allows for both the representation of different structural interpretations as well as the separation of distinct types of information about any set of events. In the CHARM specification, this is achieved through the use of *constituents* which define higher-level groupings of events without committing the user to any particular hierarchy.

Specification At the abstract level, a constituent is defined by a pair of the form $(properties, particles)$. The Particles of a constituent consist of the set of events and sub-constituents from which it is formed. A sub-constituent of a constituent is one of its particles or a sub-constituent of one of them such that no constituent may be a constituent of itself. The constituent structure of a musical piece is, therefore, a directed acyclic graph (Harris *et al.*, 1991). The properties of a constituent allow the logical specification of the structural relationship between its particles in terms of the membership of (user-defined) classes. The properties component of a constituent is a pair of the form $(spec, env)$ where $spec$ is a logical specification for the defining structural property of the constituent and env is (a possibly empty) set of values for event-like information concerning Time, Pitch etc. associated directly with the constituent.

Three very general types of constituent are given as examples by Harris *et al.* (1991): *collection* constituents, *stream* constituents and *slice* constituents. The first of these places no constraints on the structural relationships between its particles. The second, however, restricts a constituent to particles in which no particle starts between the onset and offset of any other particle (*e.g.,* a melodic phrase):

$$\begin{aligned} stream \;\Leftrightarrow\; & \forall p_1 \in particles, \neg\exists p_2 \in particles, \\ & p_1 \neq p_2 \wedge \\ & GetTime(p_1) \leq GetTime(p_2) \wedge \\ & GetTime(p_2) < add_{td}(GetTime(p_1), GetDuration(p_1)) \end{aligned}$$

The third type, on the other hand, requires that some point in time is common to every particle in the constituent:

$$\begin{aligned} slice \;\Leftrightarrow\; & \exists t \in Time, \forall p \in particles, \\ & GetTime(p) \leq t \wedge \\ & t \leq add_{td}(GetTime(p), GetDuration(p)) \end{aligned}$$

The user is free to define other kinds of constituent appropriate to the kinds of musical groupings being studied.

Implementation A constituent is defined as the tuple:

```
constituent(Identifier,Properties,Definition,
            Particles,Description)
```

where:

- the *identifier* is a unique identifier for the constituent;
- the *properties* or *structural type* defines the externally available properties of the constituent type which are derived from the externally defined interface functions;
- the *definition* or *musical type* defines the intrinsic musical properties of a constituent;
- the *particles* component contains a list of the particles of the constituent;
- the *description* is an arbitrary structure defined by the user for annotation of useful information.

As with events, each constituent must have associated selector functions which access the appropriate component of a constituent tuple. There must also be a function *putConstituent* analogous to the *putEvent* function. As with events, these selector functions are intended to be used with a database of constituent objects (Harris *et al.*, 1991). Smaill *et al.* (1993) argue that the small atomic set of interface functions they have defined for the pitch and time data types are sufficient to allow the construction of more complex operations for manipulating musical material in a wide range of musical tasks. Examples of such operations include the dilation of the interval structure between events, replacement of events by sub-constituents, blanking out of material and distinguishing voices in polyphonic music.

Wiggins *et al.* (1993) argue that the CHARM system scores well on both expressive completeness and structural generality. Regarding the latter, the use of constituents allows the explicit representation of any structural property of music at any level of abstraction and a precise characterisation of the relationships between such structures (for example, multiple views of any musical structure may be easily defined). Furthermore, the use of abstract data types facilitates the construction of functions for manipulating these musical

structures. In terms of expressive completeness, the abstraction away from implementational detail resulting from the use of abstract data types frees the system from any particular style, tonal system, tradition or application as long as the mathematical specifications are followed. Finally, this abstraction from detail also facilitates the common use of a general and expressive representation scheme for many different applications.

5.2.3 Multiple Viewpoint Representations of Music

While CHARM emphasises the internal structure of Lewin's GISs and introduces constituents for the flexible representation of groupings of events, other representation schemes emphasise Lewin's use of quotient and product GISs to allow the observation of a musical object from *multiple viewpoints* (Conklin & Cleary, 1988; Ebcioğlu, 1988). In this section, we review the representation language of the multiple viewpoint framework as developed by Conklin (1990, see also Conklin & Witten, 1995). The specific motivation in the development of the framework was to extend the application of statistical modelling techniques to domains, such as music, where events have an internal structure and are richly representable in languages other than the basic event language. Here, the framework is discussed only insofar as it applies to monophonic music; see Conklin (2002) for extensions to accommodate the representation of homophonic and polyphonic music.

Like CHARM, the multiple viewpoints framework takes as its musical surface sequences of musical events which roughly correspond to individual notes as notated in a score. Each event consists of a finite set of descriptive variables or *basic attributes* each of which may assume a value drawn from some finite domain or alphabet. Each attribute describes an abstract property of events and is associated with a type, τ, which specifies the properties of that attribute (see Table 5.1). Each type is associated with a syntactic domain, $[\![\tau]\!]$, denoting the set of all syntactically valid elements of that type. Each type is also supplied with an informal semantics by means of an associated semantic domain, $[\![[\![\tau]\!]]\!]$, which denotes set of possible meanings for elements of τ and a function, $[\![\cdot]\!]_\tau$: $[\![\tau]\!] \rightarrow [\![[\![\tau]\!]]\!]$, which returns the semantic interpretation of any element of type τ. The Cartesian product of the domains of n basic types $\tau_1, \ldots, \tau_n$ is referred to as the *event space*, ξ:

$$\xi = [\![\tau_1]\!] \times [\![\tau_2]\!] \times \ldots \times [\![\tau_n]\!]$$

An *event* $e \in \xi$ is an instantiation of the attributes $\tau_1, \ldots, \tau_n$ and consists of an

Symbol	Interpretation	Example
τ	A typed attribute	`cpitch`
$[\tau]$	Syntactic domain of τ	$\{60, \ldots, 72\}$
$\langle\tau\rangle$	Type set of τ	{`cpitch`}
$[\![\tau]\!]$	Semantic domain of τ	$\{C_4, C\sharp_4, \ldots, B_4, C_5\}$
$[\![.]\!]_\tau : [\tau] \rightarrow [\![\tau]\!]$	Semantic interpretation of $[\tau]$	$[\![60]\!]_{\texttt{cpitch}} = C_4$
$\Psi_\tau : \xi^* \rightharpoonup [\tau]$	see text	see text

Table 5.1: Sets and functions associated with typed attributes.

n-tuple in the event space. The event space ξ, therefore, denotes the set of all representable events and its cardinality, $|\xi|$, will be infinite if one or more of the attribute domains $[\tau_1]$,...,$[\tau_n]$ is infinite. Attribute types appear here in typewriter font in order to distinguish them from ordinary text.

A *viewpoint* modelling a type τ is a partial function, $\Psi_\tau : \xi^* \rightharpoonup [\tau]$, which maps sequences of events onto elements of type τ.[1] Each viewpoint is associated with a *type set* $\langle\tau\rangle \subseteq \{\tau_1, \ldots, \tau_n\}$, stating which basic types the viewpoint is derived from and is, therefore, capable of predicting (Conklin, 1990). For ease of exposition, a viewpoint will sometimes be referred to by the type it models. A collection of viewpoints forms a *multiple viewpoint system*. The nature of several distinct *classes* of viewpoint is now defined.

Basic Viewpoints For *basic types*, those associated with basic attribute domains, Ψ_τ is simply a projection function (Conklin, 1990) and $\langle\tau\rangle$ is a singleton set containing the basic type itself. An example of a basic type is one that represents the chromatic pitch of an event in terms of MIDI note numbers (`cpitch`; see Table 5.1).

Derived Viewpoints A type that does not feature in the event space but which is derived from one or more basic types is called a *derived type*. The function Ψ_τ acts as a *selector* function for events, returning the appropriate attribute value when supplied with an event sequence (Conklin, 1990). Since the function is partial the result may be undefined (denoted by $\perp$) for a given event sequence. Many of the derived types implemented by Conklin (1990) are inspired by the construction of quotient GISs developed by Lewin (1987) and reviewed in §5.2.1. The motivation for constructing such types is to capture and model

[1] While viewpoints were defined by Conklin & Witten (1995) to additionally comprise a statistical model of sequences in $[\tau]^*$, here we consider viewpoints to be a purely representational formalism and maintain a clear distinction between our representation language and our modelling strategies.

the rich variety of relational and descriptive terms in a musical language (Conklin, 1990). A viewpoint modelling a derived type is called a *derived* viewpoint and the types from which it is derived, and which it is capable of predicting, are given by the type set for that viewpoint. An example of a derived type is one which represents melodic intervals in the chromatic pitch domain (see GIS 2.1.2 discussed in §5.2.1). Given the basic type `cpitch` shown in Table 5.1, the derived viewpoint `cpint` (Conklin, 1990) is defined by the function:

$$\Psi_{\texttt{cpint}}(e_1^j) = \begin{cases} \perp & \text{if } j = 1, \\ \Psi_{\texttt{cpitch}}(e_1^j) - \Psi_{\texttt{cpitch}}(e_1^{j-1}) & \text{otherwise.} \end{cases} \tag{5.3}$$

Linked Viewpoints A system of viewpoints modelling primitive types will have limited representational and predictive power due to its inability to represent any interactions between those individual types (Conklin & Witten, 1995). *Linked viewpoints* are an attempt to address this problem and were motivated by the direct product GISs described by Lewin (1987) and reviewed in §5.2.1. A *product type* $\tau = \tau_1 \otimes \ldots \otimes \tau_n$ between *n* constituent types $\tau_1, \ldots, \tau_n$ has the following properties:

$$\begin{aligned} [\tau] &= [\tau_1] \times \ldots \times [\tau_n] \\ \langle \tau \rangle &= \bigcup_{k=1}^{n} \langle \tau_k \rangle \\ [\![\tau]\!] &= [\![\tau_1]\!] \; and \; \ldots \; and \; [\![\tau_n]\!] \\ \Psi_\tau(e_1^j) &= \begin{cases} \perp & \text{if } \Psi_{\tau_i}(e_1^j) \text{ is undefined for any } i \in \{1, \ldots, n\} \\ \Psi_{\tau_1}(e_1^j), \ldots, \Psi_{\tau_n}(e_1^j) & \text{otherwise.} \end{cases} \end{aligned}$$

A linked viewpoint is one which models a product type. Linked viewpoints add to the representation language the ability to represent disjunctions of conjunctions of attribute values (as opposed to simple disjunctions of attribute values). To give an example, it was found by Conklin & Witten (1995) that a viewpoint linking melodic pitch interval with inter-onset interval (`cpint`$\otimes$`ioi`) proved useful in modelling the chorale melodies harmonised by J. S. Bach. This finding suggests that these two attribute types are correlated in that corpus.

Test Viewpoints A *test viewpoint* models a Boolean-valued attribute type and is used to define locations in a sequence of events (Conklin & Anagnostopoulou, 2001) specifically those which are used in the construction of threaded view-

points (as discussed below). The name derives from the fact that these types perform a Boolean-valued test at a given event location. An example is the `fib` viewpoint defined by Conklin (1990) as follows:

$$\Psi_{\texttt{fib}}(e_1^j) = \begin{cases} \text{T} & \text{if } \Psi_{\texttt{posinbar}}(e_1^j) = 1, \\ \text{F} & \text{otherwise} \end{cases} \tag{5.4}$$

where `posinbar` is a derived type giving the relative position of an event in the bar (*e.g.,* $[\![1]\!]_{\texttt{posinbar}}$ = the first event in the current bar). Figure 7.2 illustrates the representation of a melodic fragment in terms of the `fib` attribute.

Threaded Viewpoints Types whose values are only defined at certain points in a piece of music (*e.g.,* the first event in each bar) are called *threaded types* and viewpoints modelling these types are called *threaded viewpoints*. Threaded viewpoints model the value of a *base viewpoint* at temporal or metric locations where a specified test viewpoint returns true and are undefined otherwise (Conklin & Anagnostopoulou, 2001). The base viewpoint may be any primitive or linked viewpoint. Threaded viewpoints were developed to take advantage of structure emerging from metrical grouping and phrasing in music. The syntactic domain of a threaded viewpoint is the Cartesian product of the domains of the base viewpoint and a viewpoint, `ioi`, representing inter-onset intervals (Conklin & Anagnostopoulou, 2001). The latter component of a threaded viewpoint element represents the *timescale* of the element: the inter-onset interval between that element and its (possibly) non-adjacent predecessor. A *periodic* threaded type threads a sequence at periods of a fixed number of events; most useful threaded types, however, will be aperiodic. To take an example, consider the `thrbar` viewpoint which is constructed from the base viewpoint `cpint` and the test viewpoint `fib` (Conklin & Witten, 1995). This viewpoint represents melodic intervals between the first events in each consecutive bar and is undefined at all other locations in a melodic sequence. Its viewpoint elements consist of pairs of `cpint` and `ioi` elements corresponding to the pitch interval between the first events in two successive bars and the inter-onset interval between those events.

It is clear from the above that any sequence of musical events can be viewed as a set of derived sequences – one for each primitive type (*i.e.,* all but product types) used. This set of sequences is represented in a *solution array* (Ebcioğlu, 1988). For n primitive viewpoints $\tau_1, \ldots, \tau_n$ and a basic event sequence e_1^j, the solution array is an $n \times j$ matrix where location (k, l) holds the value $\Psi_{\tau_k}(e_1^l)$ if it

is defined or else $\perp$ (Conklin & Witten, 1995). Product types do not need to be represented explicitly in the solution array since they can be derived from their constituent rows. For a system of *n* primitive types, 2^n distinct multiple viewpoint systems can be formed while this increases to n^n once linked viewpoints with any number of constituents are allowed (Conklin & Witten, 1995). Given this exponential relationship between the number of primitive viewpoints and the space of possible viewpoint systems, multiple viewpoint systems have typically been hand-constructed through the use of expert domain-specific knowledge to define a restricted set of basic, derived, linked, test and threaded types which are expected to be useful in modelling a given musical genre (Conklin, 1990; Conklin & Witten, 1995).

5.3 The Musical Surface

As discussed in §4.2, the electronic format in which the selected musical corpora is encoded should contain all the required information. All the datasets used in the present research were originally encoded in the `**kern` format. This section introduces the basic event space making up the musical surface assumed in the present research as well as the preprocessing of the original data into this basic representation scheme.

The `**kern` representation format is one of several in the *humdrum syntax* (Huron, 1997). It is designed to encode the syntactic information conveyed by a musical score (as opposed to orthographic information, on the one hand, and performance information on the other) for analytic purposes. Consequently, the `**kern` scheme allows encoding of pitch (*e.g.,* concert pitch, accidentals, clefs, key signatures, harmonics, glissandi and so on), duration (*e.g.,* canonic musical duration, rests, augmentation dots, grace notes, time signature and tempo), articulation (*e.g.,* fermata, trills, accents), timbre (*e.g.,* instrument and instrument class) and many other structural components of a score (*e.g.,* phrase markings, bar lines, repetitions, bowing information, beaming and stem direction). Appendix B presents an example of a `**kern` file from the EFSC (see Chapter 4).

The original data were preprocessed into an event based format similar to those used by the CHARM and multiple viewpoints frameworks (see §5.2). Since the data used in this research is purely monophonic, all compositions are represented as CHARM stream constituents. The preprocessed data were used to construct a CHARM-compliant relational database in which event attributes, events, compositions and datasets are associated with unique identifiers and

selector functions. In other respects, the representation scheme is closely based on the multiple viewpoints framework (see §5.2.3).

The basic event space ξ of the preprocessed data is the Cartesian product of the domains of nine basic types (each of which is specified in full below):

$$\begin{aligned}[\texttt{onset}] \;&\times\; [\texttt{deltast}] \times [\texttt{dur}] \times [\texttt{barlength}] \times [\texttt{pulses}] \\ &\times\; [\texttt{cpitch}] \times [\texttt{keysig}] \times [\texttt{mode}] \\ &\times\; [\texttt{phrase}]\end{aligned}$$

An event is represented as an instantiation of the component attribute dimensions of ξ. Attribute types are atomic and lack the explicit internal structure of CHARM attributes. It was felt that this increased the flexibility of attribute types, in keeping with the multiple viewpoints approach. These basic attribute types are summarised in the upper section of Table 5.2 which shows for each type τ, an informal description of its semantic interpretation function $[\![\cdot]\!]_\tau$, the syntactic domain $[\tau]$ and the type set $\langle\tau\rangle$. Note that the syntactic domains given for each attribute type are theoretical. For a given attribute type, the syntactic domain used is typically a subset of that shown in Table 5.2. In particular, the domains of basic types are generated through simple analysis of the basic elements actually occurring in the datasets involved in each experiment. The basic attribute types are described in detail below. Appendix B shows a melody from the EFSC represented both in standard music notation and as viewpoint sequences for each of the attribute types making up the basic event space used in the present research.

The onset time of an event is represented by the attribute type `onset`. The domain of onset values `[onset]` is $\mathbb{Z}^*$, the set of non-negative integers. The user may define the granularity of the time representation by setting the *timebase* during preprocessing to any appropriate positive integer. The timebase corresponds to the number of time units in a semibreve thereby limiting the granularity of the time representation to a minimum unit inter-onset interval that may be represented. Some example timebases are shown in Table 5.3 with their associated granularities. Since both demisemiquaver and semiquaver triplet durations occur in the datasets all preprocessing was carried out using a timebase of 96 (the LCM of 24 and 32). Following Conklin (1990, p. 83), the first time point of any composition is zero corresponding to the beginning of the first bar *whether complete or incomplete*. Thus, the first event in a composition may have a non-zero onset due to an opening anacrusis as in the case of the

τ	$[\![\cdot]\!]_\tau$	$[\tau]$	$\langle\tau\rangle$
onset	event onset time	$\mathbb{Z}^*$	{onset}
deltast	rest duration	$\mathbb{Z}^*$	{deltast}
dur	event duration	$\mathbb{Z}^+$	{dur}
barlength	bar length	$\mathbb{Z}^*$	{barlength}
pulses	metric pulses	$\mathbb{Z}^*$	{pulses}
cpitch	chromatic pitch	$\mathbb{Z}$	{cpitch}
keysig	key signature	{-7,-6,...,6,7}	{keysig}
mode	mode	{0,9}	{mode}
phrase	phrasing	{-1,0,1}	{phrase}
cpitch-class	pitch class	{0,...,11}	{cpitch}
cpint	pitch interval	$\mathbb{Z}$	{cpitch}
cpcint	pitch class interval	{0,...,11}	{cpitch}
contour	pitch contour	{-1,0,1}	{cpitch}
referent	referent or tonic	{0,...,11}	{keysig}
inscale	(not) in scale	{T,F}	{cpitch}
cpintfref	cpint from tonic	{0,...,11}	{cpitch}
cpintfip	cpint from first in piece	[cpint]	{cpitch}
cpintfib	cpint from first in bar	[cpint]	{cpitch}
cpintfiph	cpint from first in phrase	[cpint]	{cpitch}
posinbar	event position in bar	$\mathbb{Z}^*$	{onset}
ioi	inter-onset interval	$\mathbb{Z}^+$	{onset}
dur-ratio	duration ratio	$\mathbb{Q}^+$	{dur}
tactus	(not) on tactus pulse	{T,F}	{onset}
fib	(not) first in bar	{T,F}	{onset}
fiph	(not) first in phrase	{T,F}	{phrase}
liph	(not) last in phrase	{T,F}	{phrase}
phraselength	length of phrase	$\mathbb{Z}^+$	{phrase,onset}
thrtactus	cpint at metric pulses	[cpint] $\times \mathbb{Z}^+$	{cpitch,onset}
thrbar	cpint at first in bar	[cpint] $\times \mathbb{Z}^+$	{cpitch,onset}
thrfiph	cpint at first in phrase	[cpint] $\times \mathbb{Z}^+$	{cpitch,onset}
thrliph	cpint at last in phrase	[cpint] $\times \mathbb{Z}^+$	{cpitch,onset}

Table 5.2: The basic, derived, test and threaded attribute types used in the present research.

melody shown in Figure B.1 whose first event is a crotchet anacrusis with an onset time of 48 since the melody is in 3/4 metre.

Rests are not explicitly encoded, which means that the *inter-onset interval* between two events may be longer than the duration of the first of these events (Conklin, 1990). The temporal interval, in terms of basic time units, between the end of one event and the onset of the next (*i.e.,* a rest) is represented by the attribute deltast where $[\![0]\!]_{\texttt{deltast}}$ = no rest preceding an event. As an example of this attribute, since the melody shown in Figure B.1 contains no rests, the deltast attribute is zero for all events. While [onset] is potentially infinite,

Timebase	Granularity
1	Semibreve
2	Minim
4	Crotchet
6	Crotchet triplet
8	Quaver
12	Quaver triplet
16	Semiquaver
24	Semiquaver triplet
32	Demisemiquaver

Table 5.3: Example timebases and their associated granularities.

[deltast] is not. Following Conklin (1990), instead of placing an arbitrary bound on [onset], onset is modelled indirectly using deltast which assumes a finite domain corresponding to the set of deltast values occurring in the corpus. The duration of an event is represented in terms of basic time units by the attribute dur. The melody shown in Figure B.1 provides a clear example of the representation of event duration with its alternating pattern of crotchets and minims.

Since these attributes are defined in terms of basic time units, they are dependent on the chosen timebase. For example, with a timebase of 96, $[\![24]\!]_{\texttt{dur}}$ = crotchet, while with a timebase of 48, $[\![24]\!]_{\texttt{dur}}$ = minim. As another example, with a timebase of 96, $[\![12]\!]_{\texttt{deltast}}$ = quaver, indicating that an event is followed by a quaver rest. For the datasets used in the present research, [[dur]] ranges from a demisemiquaver to a breve while [[deltast]] ranges from a semiquaver rest to the combined duration of adjoining semibreve and dotted minim rests. Note that tied notes are collapsed during preprocessing into a single event whose onset corresponds to the onset of the first note of the tie and whose duration corresponds to the sum of the durations of the notes marked as tied.

Time signatures are represented in terms of two event attributes. The attribute barlength is a non-negative integer representing the number of time units in a bar. As an example, Conklin (1990) assumed a timebase of 16, with the result that $[\![16]\!]_{\texttt{barlength}}$ = 4/4 time signature and $[\![12]\!]_{\texttt{barlength}}$ = 3/4 time signature. The 100 chorales studied by Conklin (1990) only contained these two time signatures. In the case of the datasets used for the present research, the situation is more complicated due to the presence of compound metres. For example, given a timebase of 96 as used in this research, $[\![72]\!]_{\texttt{barlength}}$ could

indicate either 3/4 or 6/8 time. As a result, the attribute type `pulses`, which is a non-negative integer derived from the numerator of the time signature, is used to represent the number of metric pulses in a bar. For example, since the melody shown in Figure B.1 is in 3/4 metre for its entire length, the values of the `barlength` and `pulses` attributes are 78 and 3 for all events in the melody. A product type `pulses`$\otimes$`barlength` could be used to represent the time signature of a given score. Note that either of these attributes may assume a value of zero if the time signature is unspecified in the `**kern` representation (a `*MX` token) although this eventuality never arises in the selected datasets.

The *chromatic pitch* of an event is represented as an integer by the event attribute `cpitch`. The mapping from concert pitch in `**kern` to `cpitch` is defined such that `cpitch` conforms to the MIDI standard (Rothstein, 1992), *i.e.,* $[\![60]\!]_{\texttt{cpitch}}$ = C_4 or middle C. In the datasets as a whole, [`cpitch`] = {47,48,...,90,91} which means that [[`cpitch`]] ranges from B_2 to G_6. Table 4.1 shows the cardinality of [`cpitch`] individually for each of the datasets used.

Key signatures are represented by the attribute type `keysig` which may assume values in the set $\{-7, -6, \ldots, 0, \ldots, 6, 7\}$ (following Conklin, 1990, p. 84) and represents the key signature in terms of number of sharps or flats as follows:

$$\texttt{keysig} \;=\; \begin{cases} sharps & \text{if } sharps > 0 \\ -flats & \text{if } flats > 0 \\ 0 & \text{otherwise} \end{cases}$$

In the datasets used in the present research, [`keysig`] = {-5,-4,...,3,4} where, for example, $[\![\texttt{-5}]\!]_{\texttt{keysig}}$ = 5 flats, $[\![\texttt{4}]\!]_{\texttt{keysig}}$ = 4 sharps and $[\![\texttt{0}]\!]_{\texttt{keysig}}$ = no sharps or flats. The mode of a piece is represented by the event attribute `mode` where, in theory, [`mode`] = {0,1,...,11}. In the datasets used in this research, however, [`mode`] = {0,9} where $[\![\texttt{0}]\!]_{\texttt{mode}}$ = major and $[\![\texttt{9}]\!]_{\texttt{mode}}$ = minor, reflecting the fact that the minor mode corresponds to rotation of the pitch class set corresponding to its relative major scale by 9 semitones (see Balzano, 1982). As an example, since the melody shown in Figure B.1 has a single key signature consisting of a single sharpened F, the value of the `keysig` attribute is 1 for all events. Furthermore, since the `**kern` source for this melody, shown in Appendix B, indicates (via the token `*G:`) that it is in the key of G major, the value of the `mode` attribute is 0 for all events in the melody (see Figure B.1). Although this scheme allows for the representation of the Church modes, they cannot be represented in the `**kern` format. This basic attribute is included in order to allow the calculation of the tonic which is not possible using the `keysig`

attribute in isolation and which is useful in modelling the effects of tonality (see §5.4.1).

Finally, phrase level features are represented by the event attribute `phrase` whose domain [`phrase`] = {-1,0,1} where: $[\![1]\!]_{\texttt{phrase}}$ = the event is the first in a phrase; $[\![-1]\!]_{\texttt{phrase}}$ = the event is the last in a phrase; and $[\![0]\!]_{\texttt{phrase}}$ = the event is neither the first nor the last in a phrase. The value is derived directly from the original `**kern` encodings of the EFSC, where phrases are grouped by braces (*i.e.,* `{}`) in the `**kern` format. As an illustration of this aspect of the preprocessing, consider the melody shown in Appendix B. The braces in the `**kern` source indicates that the melody consists of a single phrase. Accordingly, the `phrase` attribute assumes values of zero for all events in the melody with the exception of the first, which assumes a value of 1 indicating that it is the first event in a phrase, and the last, which assumes a value of -1 indicating that it is the last event in a phrase (see Figure B.1). The phrase markings in the EFSC were taken from the form of the text, and nested or overlapping phrases (denoted by `&{` and `&}` in the `**kern` format) do not appear in the datasets chosen for this research. Following Conklin (1990, p. 84), fermata (signified by a semi-colon in the `**kern` format) are also used as indicators for phrase endings in the case of Dataset 2 (see Figure 8.6). In such cases, events immediately following an event under a fermata are assumed to represent the start of a new phrase. The first event in a piece is considered to represent an implicit phrase beginning.

A few general points are worth noting about the preprocessing. First, repetitions of musical material are not explicitly expanded – the sections follow each other as they are encoded in the original `**kern` file. A second issue regards the representation of time signatures, key signatures and phrase boundaries. In its original formulation, the CHARM specification focused on performed musical objects and did not include such information as time signature, key signature or phrasing (see §5.2.2). However, since the musical works to be represented in the current research are composed rather than performed objects, time signature (`barlength`), key signature (`keysig`) and phrase boundaries (`phrase`) are explicitly represented. Conklin (1990, p. 82) notes that there are two alternative possibilities for representing such features: first, to prefix complete sequences with appropriate identifiers; and second, to include them as event attributes (as shown in Figure B.1). Conklin argues that the latter approach is to be preferred on the grounds that it makes for a parsimonious multiple viewpoint framework:

- it ensures that no special circumstances have to be provided for the pre-

diction of these features;

- it allows these attributes to be linked with other attributes of event sequences;
- it allows for the encoding of compositions which change key or time signature.

For these reasons, time signature, key signature, mode and phrase boundaries are represented as attributes of events. Finally, while other attributes of events, such as tempo and dynamics, can be represented in `**kern` and other Humdrum representations, these do not appear in the datasets used and were not included.[2]

5.4 The Multiple Viewpoint Representation

In addition to the basic viewpoints described in §5.3, a number of derived, test, threaded and linked viewpoints have been developed in the present research. These viewpoints are primarily motivated by concerns with modelling pitch structure (see §1.4). The perception of pitch itself can only be explained using a multidimensional representation in which a number of perceived equivalence relations are honoured (Shepard, 1982). In the perception of melodies, more generally, research has demonstrated that pitch is only one of several interacting musical dimensions that impinge upon our perceptions (Balzano, 1986a; Boltz, 1993; Schmuckler & Boltz, 1994; Tekman, 1997; Thompson, 1993, 1994).

The approach adopted here for capturing such phenomena has been to implement a handful of viewpoints (see §5.2.3) corresponding to melodic dimensions that are considered to be relevant on the basis of previous research in music perception and the computational modelling of music. Note, however, that the implemented viewpoints are not intended to represent an exhaustive coverage of the space of possible psychological and computational hypotheses regarding the representation of music. For example, it might be expected that a derived viewpoint modelling inter-onset interval ratio would be more relevant to the examination of melody perception in Chapter 8 than a viewpoint modelling duration ratio (`dur-ratio`, see §5.4.1) although this hypothesis has

[2]Tempo, for example, is encoded as crotchet beats per minute in the `**kern` format and is signified by the token `*MM`. A notable limitation of the `**kern` format is its inability to represent musical dynamics. Several other Humdrum representations permit the representation of dynamics (*e.g.,* the `**dyn`, `**dynam` and `**db` representations).

not been examined in the present research. The possibility of automatically constructing derived, threaded and linked viewpoints on the basis of objective criteria is discussed in §10.3.

The derived, test, threaded and product types implemented and used in the present research – many of which are inspired by those developed by Conklin & Witten (1995) – are described and justified in this section. Table 5.2 contains a summary of these types, showing, respectively, derived, test and threaded types in its second, third and fourth vertical sections. Note that not all of these types will be useful for modelling pitch structure: some simplify the expression of more useful types while others will only be useful when linked with attribute types derived from `cpitch`. The product types used in the present research are summarised in Table 5.4 and discussed in detail in §5.4.4.

5.4.1 Derived Types

As noted above, studies of pitch perception alone reveal evidence of the representation of multiple interacting dimensions. In particular, the perceived relation between two pitches reflects the interaction of a number of different equivalence relations, of which the most influential is octave equivalence (Balzano, 1982; Krumhansl, 1979; Krumhansl & Shepard, 1979; Shepard, 1982). For example, Krumhansl & Shepard (1979) carried out a probe tone experiment (see §3.6) in which the contexts consisted of the first seven notes of an ascending or descending major scale and the probe tones were selected from the set of 13 diatonic pitches in the range of an octave above and below the first tone of the context. The results revealed that all subjects gave the highest rating to the tonic, its octave neighbour and the tones belonging to the scale while an effect of pitch height was only found in the case of the least musical subjects. In fact, the octave appears to be "a particularly privileged interval" in the musics of most cultures (Sloboda, 1985, p. 254). In view of these findings, a derived type `cpitch-class` was created to represent pitch class or chroma in which, for example, C_4 is equivalent to C_3 and C_5. As described in §5.2.1, this type is constructed by applying the following congruence relation to `cpitch` (Conklin, 1990; Lewin, 1987):

$$i \equiv j \longleftrightarrow (i - j) \bmod 12 = 0$$

from which it is evident that [`cpitch-class`] = {0,1,...,10,11}.

In addition to the psychological existence of equivalence relations on pitch itself, there is also evidence that the interval structure of melodies is encoded,

retained and used in recognition memory for melodies (Dowling, 1978; Dowling & Bartlett, 1981). For example, Dowling & Bartlett (1981) conducted experiments in which subjects listened to a melody and, after an interval, were asked to detect copies of input melodies (targets) as well as related items which replicated the contour and rhythm of the target but differed in terms of pitch interval. The results demonstrated substantial retention of pitch interval information over retention periods of several minutes. On the basis of evidence such as this, a derived type `cpint` was developed to represent chromatic pitch interval. The viewpoint $\Psi_{\texttt{cpint}}()$ is defined in Equation 5.3 and $[\texttt{cpint}] = \mathbb{Z}$. From a computational perspective, modelling pitch interval enables representation of the equivalence of melodic structures under transposition. Figure 7.2 illustrates the representation of a melodic fragment in terms of the `cpint` attribute.

Two further derived types were developed in order to represent more abstract properties of pitch intervals. The first, `cpcint` represents octave equivalent pitch class interval and its derivation from `cpint` is similar to that of `cpitch-class` from `cpitch`. Thus, for example, $[\![0]\!]_{\texttt{cpcint}}$ = unison, $[\![4]\!]_{\texttt{cpcint}}$ = major third, $[\![7]\!]_{\texttt{cpcint}}$ = perfect fifth and so on. The perceptual motivations for using `cpcint` are similar to those for `cpitch-class`. The second derived type, `contour` is an even more abstract type representing pitch contour where:

$$\Psi_{\texttt{contour}}(e_1^j) = \begin{cases} -1 & \text{if } \Psi_{\texttt{cpitch}}(e_1^j) < \Psi_{\texttt{cpitch}}(e_1^{j-1}) \\ 0 & \text{if } \Psi_{\texttt{cpitch}}(e_1^j) = \Psi_{\texttt{cpitch}}(e_1^{j-1}) \\ 1 & \text{if } \Psi_{\texttt{cpitch}}(e_1^j) > \Psi_{\texttt{cpitch}}(e_1^{j-1}) \end{cases}$$

It is evident from this definition that $[\texttt{contour}] = \{\text{-}1,0,1\}$ where: $[\![1]\!]_{\texttt{contour}}$ = ascending interval; $[\![0]\!]_{\texttt{contour}}$ = unison; and $[\![\text{-}1]\!]_{\texttt{contour}}$ = descending interval. It has been demonstrated that listeners are highly sensitive to contour information in recognition memory for melodies (Deutsch, 1982; Dowling, 1978, 1994). Furthermore, it has proved fruitful to represent pitch class interval and pitch contour in research on musical pattern matching and discovery (Cambouropoulos, 1996; Conklin & Anagnostopoulou, 2001).

Scale degree (or diatonic pitch) is a highly influential property of music both at the perceptual level (Balzano, 1982; Krumhansl, 1979) and at the analytical level (Cambouropoulos, 1996; Meredith, 2003). However, derived types for diatonic pitch name and accidentals were not included in the present research due to the fact that while chromatic pitch is easily and reliably derivable from the pitch names used in traditional Western staff notation, the converse

is not true. The development of reliable style-independent algorithms for pitch spelling is the subject of ongoing research (see, *e.g.,* Meredith, 2003).

Another set of derived types included in the present research were designed to represent relative pitch structure and, in particular, structures defined and perceived in relation to an induced tonal centre. For our purposes, tonality induction refers to the process by which a listener infers a tonal reference pitch (the tonal centre) and perceives other tones in relation to this pitch (Vos, 2000). Research has demonstrated that induced tonality has a significant impact on such aspects of music perception as recognition memory for melodies (Cohen *et al.*, 1977; Dowling, 1978) and the ratings of tones in key-defining contexts (Krumhansl, 1979; Krumhansl & Shepard, 1979). In general, there is a wealth of evidence that listeners implicitly induce a tonality which guides their expectations for and interpretations of subsequent musical structures (Krumhansl, 1990).

In order to model the effects of tonality, a derived type called `referent` has been developed which represents the referent or tonic at a given moment in a melody. This type is derived from the basic type `keysig` and uses the basic type `mode` to disambiguate relative major and minor keys. The viewpoint for `referent` is defined as follows:

$$\Psi_{\texttt{referent}}(e_1^j) = \Psi_{\texttt{mode}}(e_1^j) + \begin{cases} (\Psi_{\texttt{keysig}}(e_1^j) \times 7) \bmod 12 & \text{if } \Psi_{\texttt{keysig}}(e_1^j) > 0 \\ (\Psi_{\texttt{keysig}}(e_1^j) \times -5) \bmod 12 & \text{if } \Psi_{\texttt{keysig}}(e_1^j) < 0 \\ 0 & \text{otherwise} \end{cases}$$

assuming that middle C is represented by an integer multiple of 12 (*e.g.,* 0 12 24 36 48 60 and so on).

While the `referent` type is not, in and of itself, very useful for modelling pitch, it allows the derivation of other types which are relevant to modelling the influences of tonality on pitch. The derived type `cpintfref`, for example, represents the pitch interval of a given event from the tonic. The domain of this type, [`cpintfref`] = {0,1,...,10,11} where, for example, $[\![0]\!]_{\texttt{cpintfref}}$ = tonic, $[\![4]\!]_{\texttt{cpintfref}}$ = mediant, $[\![7]\!]_{\texttt{cpintfref}}$ = dominant and so on. This viewpoint is motivated by the hypothesis that melodic structure is influenced by regularities in pitch defined in relation to the tonic. Figure 7.2 illustrates the representation of a melodic fragment in terms of the `cpintfref` attribute. The `referent` attribute type also allows the derivation of the Boolean valued type `inscale` (Conklin & Witten, 1995) which represents whether or not an event is in the appropriate major or natural minor diatonic scale constructed on the referent.

Note that `inscale` is not classified as a test type since it has not been used to construct any threaded types in the present research (see §5.2.3).

For a listener, identifying a tonal centre may, in turn, involve identification of the most important or salient pitch or pitches based on information about duration, repetition, intensity, induced metric structure and prominence in terms of primacy or recency (Cohen, 2000). While tonality induction has been a central topic in music perception, research has tended to focus on melodic and harmonic influences with other factors (including metre and rhythm) receiving relatively little attention (Krumhansl, 1990; Vos, 2000). However, with short melodic contexts, such as those used in the current research, it is quite possible that the induced tonality will reflect the influence of pitches other than the actual tonic which are salient for a number of other reasons. In order to accommodate the representation of melodic structures in relation to other influences on pitch salience, a number of derived types have been developed which are analogous to `cpintfref`. The first of these, `cpintfip` represents the pitch interval of an event from the first event in the piece. The motivation for using this type is to capture the effects of *primacy* on perceptual and structural salience (Cohen, 2000). These effects are exploited in some computational models of tonality induction (Longuet-Higgins & Steedman, 1971).[3] The second type is `cpintfib` which represents the pitch interval of an event from the first event in the current bar. The motivation for using this type is to capture the effects of metric salience on relative pitch structure (see also §5.4.3). The final type is `cpintfiph` which represents the pitch interval of an event from the first event in the current phrase. The motivation for using this type is to capture the effects of phrase level salience on relative pitch structure (see also §5.4.3).

Another set of derived types was developed to represent the temporal organisation of melodies. Lee (1991) argues that musical events may be temporally organised in at least two different ways: first, using grouping structure; and second, using metrical structure. The first of these describes how events are grouped into various kinds of perceptual unit such as motifs, phrases, sections, movements and so on. The second concerns the manner in which a listener arrives "at a particular interpretation of a sequence, given that any sequence could be the realisation of an indefinite number of metrical structures" (Lee, 1991, p. 62).

[3] In the context of metre induction, Lee (1991) has demonstrated that listeners prefer metrical interpretations in which the pulse begins on the first event of a rhythmic sequence and are subsequently reluctant to revise this interpretation (although they may do so in the interest of obtaining a pulse at some level of metrical structure whose interval is in a preferred range of 300 to 600ms).

Regarding the perception of metrical structure, Povel & Okkerman (1981) have shown that in the absence of any other accent information (*e.g.,* dynamics) relatively long events and those initiating a cluster of events are perceived as salient. Povel & Essens (1985) demonstrated that patterns in which these perceptually salient events coincide with the beats of a particular metre are learnt more quickly and are more accurately reproduced than patterns for which this is not the case.

In an experiment which demonstrated that sequences which do not strongly induce a clock (those whose perceived accents do not coincide with the beats of any metre) are harder for subjects to reproduce, Povel & Essens (1985) found that the trend reached a plateau beyond a certain point. They propose that clock induction in these sequences is so weak that no internal clock is induced at all. Instead they argue that in these cases the subjects had to rely on grouping strategies to find metrical structure in the sequences. The comments of the subjects seemed to support this conjecture as do the results of fMRI studies (*e.g.,* Brochard *et al.*, 2000) and neuropsychological research (*e.g.,* Liegeoise-Chauvel *et al.*, 1998; Peretz, 1990).

Lerdahl & Jackendoff (1983) have suggested a number of preference rules which they claim characterise the manner in which humans establish local grouping boundaries through the perception of salient distinctive transitions at the musical surface (see also §3.3). The boundaries between groups are defined by the existence of perceived accents in the musical surface and, according to the theory, the existence of such accents depends on two principles: event proximity and event similarity. Examples of the former principle include rules which state that segmentation occurs at the end of slurs or rests and after a prolonged sound amongst shorter ones. Examples of the latter principle include rules which state that segmentation occurs at large pitch intervals and large changes in the length of events. Deliège (1987) has found empirical evidence that both musicians and non-musicians use these principles in segmenting the musical surface into groups.

In order to represent the influence of rhythmic accents on metrical and grouping structure, a number of attribute types have been derived from `onset` and `dur`. In addition to the basic types `dur` and `deltast`, the derived type `ioi` represents absolute time intervals, specifically the inter-onset interval between an event and its predecessor. The viewpoint for `ioi` is defined as follows:

$$\Psi_{\texttt{ioi}}(e_1^j) = \begin{cases} \perp & \text{if } j = 1 \\ \Psi_{\texttt{onset}}(e_1^j) - \Psi_{\texttt{onset}}(e_1^{j-1}) & \text{otherwise} \end{cases}$$

This derived type is similar to GIS 2.2.1 described by Lewin (1987) which was used as an example in §5.2.1. Figure 7.2 illustrates the representation of a melodic fragment in terms of the `ioi` attribute.

As noted above, perceived rhythmic accents may also be determined by relative, as well as absolute, time intervals (*e.g.,* durations). In order to model relative durations, the derived type `dur-ratio`, representing the ratio of the duration of one event to that of the event preceding it, was introduced. The viewpoint for `dur-ratio` is defined as follows:

$$\Psi_{\texttt{dur-ratio}}(e_1^j) = \begin{cases} \perp & \text{if } j = 1 \\ \frac{\Psi_{\texttt{dur}}(e_1^j)}{\Psi_{\texttt{dur}}(e_1^{j-1})} & \text{otherwise} \end{cases}$$

Accordingly, $[\![2]\!]_{\texttt{dur-ratio}}$ indicates that the duration of an event is twice that of its predecessor and $[\![\frac{1}{2}]\!]_{\texttt{dur-ratio}}$ indicates that the duration of an event is half that of its predecessor. Duration ratios enable the representation of equivalences of rhythmic figures under augmentation or diminution (Cambouropoulos *et al.*, 1999). This derived type is similar to GIS 2.2.3 described by Lewin (1987) which was used as an example in §5.2.1. Evidently, the types `onset`, `duration`, `deltast` and others derived from these cannot be used to predict `pitch` directly. However, such types can be used fruitfully as components in product types for representing the interaction of pitch and time based attributes in melodic structures (see §5.4.4).

Finally, the derived type `posinbar` represents the sequential position of an event in the current bar as a positive integer. Its primary function is to facilitate the expression of the test type `fib` (see §5.4.2) and the threaded type `thrbar` which is derived from `fib` (see §5.4.3).

5.4.2 Test Types

In addition to these derived types, a number of test types (see §5.2.3) have been derived from features, such as time signature and phrasing boundaries, which are explicitly encoded in the original `**kern` data. The purpose of these types is to represent events which are salient in terms of metrical or phrase structure. In §5.4.1, a number of derived types were developed to represent some of the musical primitives which determine the perceptual salience of events. The perception of temporal organisation in music, as determined by both metrical and grouping structure, depends on the perception of such events.

Metric structure in Western music is typically associated with at least two nested time periods (Jones, 1987): first, the tactus level or beat period; and

second, the bar. Palmer & Krumhansl (1990) have investigated the perceptual validity of such metric hierarchies using an extension of the probe tone methodology (see §3.6) to rhythmic context and probe tones. Subjects were asked to rate a series of probe tones inserted at various temporal locations into a metrical context based on a variety of different time signatures. The responses obtained demonstrated that listeners represent multiple temporal periodicities which are sensitive to the time signature and which coincide with music-theoretic predictions. Furthermore, the depth of the hierarchy tended to increase with musical training.

Two test types have been implemented to represent salience arising from the two most influential metrical time periods, the tactus and the bar. First, the type `tactus` represents whether or not an event occurs on a tactus pulse in the metric context defined by the current key signature. This type is derived from `onset` using `barlength` and `pulses` as follows:

$$\Psi_{\texttt{tactus}}(e_1^j) = \begin{cases} T & \text{if } \Psi_{\texttt{onset}}(e_1^j) \bmod \frac{\Psi_{\texttt{barlength}}(e_1^j)}{\Psi_{\texttt{pulses}}(e_1^j)} = 0 \\ F & \text{otherwise} \end{cases}$$

The derivation of this type illustrates the utility of representing both the numerator and denominator of the time signature as basic types. The second type, `fib` represents whether or not an event is the first event in a bar as described in §5.2.3.

Two more test types, `fiph` and `liph` (derived trivially from `phrase`) were developed to distinguish events which may be salient by virtue of, respectively, opening and closing a phrase.

5.4.3 Threaded Types

Threaded types represent the values of other types at periodic intervals defined by the value of a given test viewpoint and are elsewhere undefined (see §5.2.3). Their purpose is to allow the representation of higher level structure within the multiple viewpoints framework. Four threaded types have been developed in the present research corresponding to the four test types described in §5.4.2.

In order to represent the influence of metrical organisation on higher level pitch structure, two threaded types were created which represent pitch intervals between events occurring on tactus beats (`thrtactus`) and the first event in each bar (`thrbar`). In research using a multiple viewpoint framework for pattern discovery in Dataset 2, Conklin & Anagnostopoulou (2001) found significant repeated patterns in a threaded type similar to `thrtactus`. In addition

τ_1		τ_2
cpitch	$\otimes$	dur
cpitch	$\otimes$	ioi
cpitch	$\otimes$	dur-ratio
cpint	$\otimes$	dur
cpint	$\otimes$	ioi
cpint	$\otimes$	dur-ratio
contour	$\otimes$	dur
contour	$\otimes$	ioi
contour	$\otimes$	dur-ratio
cpintfref	$\otimes$	dur
cpintfref	$\otimes$	ioi
cpintfref	$\otimes$	dur-ratio
cpintfref	$\otimes$	fib
cpintfref	$\otimes$	cpintfip
cpintfref	$\otimes$	cpint
cpintfiph	$\otimes$	contour
cpintfib	$\otimes$	barlength

Table 5.4: The product types used in the present research.

to higher level structure defined by metric hierarchies, two threaded types were developed to represent the influence of phrasing on higher level pitch organisation. The threaded types `thrfiph` and `thrliph` model pitch intervals between, respectively, the first and last events in each successive phrase.

A threaded type represents both the value of a base viewpoint and the interval in basic time units over which that value is defined (see §5.2.3). Note that while the timescale of `thrtactus` elements will remain constant, as long as the time signature remains constant, the timescale of `thrbar` will vary depending on whether or not the events involved are preceded by rests. Likewise, the timescales of `thrliph` and `thrfiph` elements will vary depending on the number of events in the current and previous phrases. The role of the type `phraselength` is to facilitate calculating the timescale for `thrfiph` and `thrliph`.

5.4.4 Product Types

This section introduces the product types used in the present research. Given the set of basic, derived and test types shown in Table 5.2, it will be clear that the theoretical space of possible product types is very large indeed. In order to prune this space to a more manageable size, the present research only considers

a small set of links between two component types which are motivated on the basis of previous research in music perception and the computational modelling of music. Note, however, that the selected linked types are not intended to represent an exhaustive coverage of the space of possible psychological and computational hypotheses regarding the representation of music. This issue is discussed further in §10.3 which includes some recommendations for the development of the representation scheme in future research, in particular in terms of the automatic construction of derived, threaded and linked types.

As noted above, pitch is only one of several interacting influences on our perception of music. The *theory of dynamic attending* (Jones, 1981, 1982, 1987, 1990; Jones & Boltz, 1989) proposes that tonal-harmonic, melodic and temporal structure interact to dynamically guide the listener's attention to salient events as a piece of music proceeds. In particular, according to the theory, temporal accents (as discussed in §5.4.1) and melodic accents combine to yield higher order *joint accent* structure (Jones, 1987). Experimental research has confirmed that joint accent structure influences the ability to detect deviant pitch changes (Boltz, 1993), judgements of melodic completion (Boltz, 1989a), estimates of a melody's length (Boltz, 1989b), recognition memory for melodies (Jones, 1987) and the reproduction of melodies (Boltz & Jones, 1986). In addition, viewpoints linking melodic attribute types (*e.g.,* `cpint`) with rhythmic attribute types (*e.g.,* `dur`) have proved important in computational analysis, including pattern discovery (Conklin & Anagnostopoulou, 2001) and statistical prediction (Conklin & Witten, 1995), of the chorale melodies harmonised by J. S. Bach (Dataset 2).

In order to represent regularities in joint melodic and rhythmic structure, a number of product types (see §5.2.3) were constructed reflecting the conjunction of several simple dimensions of pitch structure (`cpitch`, `cpint`, `contour` and `cpintfref`) and some simple defining dimensions of rhythmic accents (`dur`, `ioi` and `dur-ratio`). As an illustration of the use of linked features to represent joint rhythmic and melodic structure, Figure 7.2 shows the representation of a melodic fragment in terms of the product type `cpint`⊗`ioi`. In order to represent regularities in joint melodic and tonal-harmonic structure, the product types `cpintfref`⊗`cpintfip` and `cpintfref`⊗`cpint` were created. These types have proved useful in previous research on the statistical prediction of melodies (Conklin, 1990; Conklin & Witten, 1995) which also motivated the inclusion of the linked types `cpintfiph`⊗`contour`, `cpintfib`⊗`barlength` (Conklin, 1990) and `cpintfref`⊗`fib` (Conklin & Witten, 1995).

5.5 Summary

In this chapter, the scheme used in the present research for the symbolic representation of music has been presented. In §5.2, a number of existing frameworks, on which the current representation scheme is based, were discussed. The preprocessing of the data (described in Chapter 4) and the basic event representation employed in this research were presented in §5.3. Finally, in §5.4 the multiple viewpoint framework developed in the current research was described in detail. The individual attribute types developed have been motivated in terms of previous research on music cognition and the computational modelling of music.

CHAPTER 6

A PREDICTIVE MODEL OF MELODIC MUSIC

6.1 Overview

In this chapter, the performance of a range of statistical models is investigated in an application-independent prediction task using a variety of monophonic music data (see §4.3). The objective is to undertake an empirical investigation of several methods for addressing the limitations of finite context grammars for modelling music (see §6.2). These methods include, in particular, a technique for combining the predictions of n-gram models called *Prediction by Partial Match* (PPM), originally developed by Cleary & Witten (1984), which forms the central component in some of the best performing data compression algorithms currently available (Bunton, 1997). Outside the realm of data compression, PPM has been used to model natural language data (Chen & Goodman, 1999) and music data (Conklin & Witten, 1995). Since its introduction, a great deal of research has focused on improving the compression performance of PPM models and the specific aim in this chapter is to evaluate the performance of these improved techniques in predicting the datasets described in §4.3.

The research contained in this chapter and Chapter 7 may be classified as basic AI in the sense discussed in §2.3. The goal is to develop powerful statistical models of melodic structure which have the potential for simulating intelligent behaviour in the context of some of the specific musical tasks cited in §3.4. The chapter is organised as follows. In §6.2, n-gram modelling is introduced and the PPM scheme is described in detail. The information-theoretic performance metrics used in the present research are also discussed. Much of the

background for the present research is drawn from the fields of statistical language modelling (Manning & Schütze, 1999) and text compression (Bell *et al.*, 1990) since research in these fields is at a more mature stage of development than in the musical domain. However, as demonstrated in this chapter, practical techniques and methodologies from these fields can be usefully applied in the modelling of music. The empirical methodology employed in the experiments is discussed in §6.3, which also contains a summary of the cross product of PPM features to be evaluated. Finally, the results of the experiments are presented in §6.4 and discussed in §6.5. The predictive systems developed in this chapter are applied to the task of modelling a single event attribute, chromatic pitch. In Chapter 7, the prediction performance of these models is examined within the multiple viewpoint framework presented in §5.4.

6.2 Background

6.2.1 Sequence Prediction and *N*-gram Models

For the purpose of describing the models developed in the present research, the acquisition of knowledge about melodic music will be characterised as a problem of learning to predict sequences (Dietterich & Michalski, 1986). The objects of interest are sequences of event attributes of a given type τ where each symbol in a given sequence e_1^j is drawn from the finite alphabet $[\tau]$ as described in Chapter 5. For the purposes of exposition we assume here a one-dimensional event space $\xi = [\tau]$. The goal of sequence learning is to derive from example sequences a model which, supplied with a sequence e_1^j, estimates the probability function $p(e_i|e_1^{i-1})$ for all $i \leq j$. It is often assumed in statistical modelling that the probability of the next event depends only on the previous $n-1$ events, for some $n \in \mathbb{Z}^+$:

$$p(e_i|e_1^{i-1}) \approx p(e_i|e_{(i-n)+1}^{i-1})$$

An example of such a model is the n-gram model introduced in §3.4 where the quantity $n-1$ represents the order of the model. Since the use of fixed order n-grams imposes assumptions about the nature of the data (see §6.2.3.6), the selection of an appropriate n is an issue when designing and building n-gram models. If the order is too high, the model will overfit the training data and fail to capture enough statistical regularity; low order models, on the other hand, suffer from being too general and failing to represent enough of the structure

present in the data. The optimal order for an n-gram model depends on the nature of the data to which it is applied and, in the absence of specific *a priori* knowledge about that data, can only be determined empirically.

An n-gram *parameter* is the probability of the prediction occurring immediately after the context. The parameters are typically estimated on some corpus of example sequences. There are several different means of estimating n-gram parameters, the simplest of which is the *Maximum Likelihood* (ML) method which estimates the parameters as:

$$p(e_i|e_{(i-n)+1}^{i-1}) = \frac{c(e_i|e_{(i-n)+1}^{i-1})}{\sum_{e\in[\tau]} c(e|e_{(i-n)+1}^{i-1})}$$

where $c(g)$ denotes the frequency count for n-gram g. In n-gram modelling, the probability of a sequence of events is expressed, following the chain rule, as the product of the estimated probabilities of the events (conditional on the identity of the previous $n-1$ events) from which it is composed:

$$p(e_1^j) = \prod_{i=1}^{j} p(e_i|e_{(i-n)+1}^{i-1}).$$

When $n > i$, at the beginning of the sequence for example, padding symbols must be introduced to provide the necessary contexts.

Fixed order ML models will run into trouble if, as a result of data sparseness, they encounter as-yet-unseen n-grams during prediction. In particular, if the model encounters a novel n-gram context or a symbol which has not previously appeared after an existing context (the zero-frequency problem – see Witten & Bell, 1991), the ML estimate will be zero. In these situations, the estimated probability of a novel n-gram will be too low and consequently the estimated probability of n-grams with non-zero counts will be too high. Additionally, the information-theoretic performance measures used in the present research (see §6.2.2) require that every symbol is predicted with non-zero probability.

In statistical language modelling, a set of techniques known collectively as *smoothing* are commonly used to address these problems. The central idea of smoothing is to adjust the ML estimates in order to generate probabilities for as-yet-unencountered n-grams. This is typically achieved by combining the distributions generated by an h-gram model with some fixed *global order bound* h with distributions less sparsely estimated from lower order n-grams (where $n < h$). Most existing smoothing techniques can be expressed using the frame-

work described in Equation 6.1 (Kneser & Ney, 1995).

$$p(e_i|e_{(i-n)+1}^{i-1}) = \begin{cases} \alpha(e_i|e_{(i-n)+1}^{i-1}) & \text{if } c(e_i|e_{(i-n)+1}^{i-1}) > 0 \\ \gamma(e_i|e_{(i-n)+1}^{i-1})p(e_i|e_{(i-n)+2}^{i-1}) & \text{if } c(e_i|e_{(i-n)+1}^{i-1}) = 0 \end{cases} \tag{6.1}$$

For a given context $e_{(i-n)+1}^{i-1}$, if a given symbol e_i occurs with a non-zero count (*i.e.,* $c(e_i|e_{(i-n)+1}^{i-1}) > 0$) then the estimate $\alpha(e_i|e_{(i-n)+1}^{i-1})$ is used; otherwise, we recursively *backoff* to a scaled version of the $(n-2)^{th}$ order estimate $p(e_i|e_{(i-n)+2}^{i-1})$ where the scaling factor $\gamma(e_i|e_{(i-n)+1}^{i-1})$ is chosen to ensure that the conditional probability distribution over the alphabet sums to unity: $\sum_{e\in[\tau]} p(e|e_{(i-n)+1}^{i-1}) = 1$. The recursion is typically terminated with the zeroth order model or by taking a uniform distribution over $[\tau]$. The various smoothing algorithms differ in terms of the techniques employed for computing $\alpha(e_i|e_{(i-n)+1}^{i-1})$ and $\gamma(e_i|e_{(i-n)+1}^{i-1})$.

An alternative to backoff smoothing is *interpolated* smoothing in which the probability of an n-gram is always estimated by recursively computing a weighted combination of the $(n-1)^{th}$ order distribution with the $(n-2)^{th}$ order distribution as described in Equation 6.2.

$$p(e_i|e_{(i-n)+1}^{i-1}) = \alpha(e_i|e_{(i-n)+1}^{i-1}) + \gamma(e_{(i-n)+1}^{i-1})p(e_i|e_{(i-n)+2}^{i-1}) \tag{6.2}$$

Detailed empirical comparisons of the performance of different smoothing techniques have been conducted on natural language corpora (Chen & Goodman, 1999; Martin *et al.*, 1999). One of the results of this work is the finding that, in general, interpolated smoothing techniques outperform their backoff counterparts. Chen & Goodman (1999) found that this performance advantage is restricted, in large part, to n-grams with low counts and suggest that the improved performance of interpolated algorithms is due to the fact that low order distributions provide valuable frequency information about such n-grams.

6.2.2 Performance Metrics

There exist many (more or less application dependent) ways of assessing the quality of an n-gram model and the ultimate evaluation metric can only be the impact it has on a specific application. Here, however, the objective is to examine the performance of such models in an application-neutral manner. It is common in the field of statistical language modelling to use information-theoretic, in particular entropy-based, measures to evaluate statistical models

of language. These metrics have been employed in the current research and they are briefly introduced below.

Given a probability mass function $p(a \in \mathcal{A}) = p(\mathcal{X} = a)$ of a random variable $\mathcal{X}$ distributed over a discrete alphabet $\mathcal{A}$ such that the individual probabilities are independent and sum to unity, *entropy* is defined according to Equation 6.3.

$$H(p) = H(\mathcal{X}) = -\sum_{a \in \mathcal{A}} p(a) \log_2 p(a) \tag{6.3}$$

Shannon's fundamental coding theorem (Shannon, 1948) states that entropy provides a lower bound on the average number of binary bits per symbol required to encode an outcome of the variable $\mathcal{X}$. The corresponding upper bound, H_{max} shown in Equation 6.4, occurs in the case where each symbol in the alphabet has an equal probability of occurring, $\forall a \in \mathcal{A}, p(a) = \frac{1}{|\mathcal{A}|}$.

$$H_{max}(p) = H_{max}(\mathcal{A}) = \log_2 |\mathcal{A}| \tag{6.4}$$

Under this interpretation, entropy is a measure of the information content of an outcome of $\mathcal{X}$ such that less probable outcomes convey more information than more probable ones. A complementary quantity, *redundancy* provides a measure of how much non-essential information is contained in an observed outcome. The redundancy R of a sequence is defined as:

$$R(p) = 1 - \frac{H(p)}{H_{max}(p)}. \tag{6.5}$$

A redundancy of zero implies maximum uncertainty and information content in an observed outcome of $\mathcal{X}$ while greater values (to a maximum of one) indicate increasing degrees of predictable information in the outcome. Entropy has an alternative interpretation in terms of the degree of uncertainty that is involved in selecting a symbol from an alphabet: greater entropy implies greater uncertainty.

In practice, the true probability distribution of a stochastic process is rarely known and it is common to use a model to approximate the probabilities expressed in Equation 6.3. *Cross entropy* is a quantity which represents the divergence between the entropy calculated from these estimated probabilities and the source entropy. Given a model which assigns a probability of $p_m(a_1^j)$

to a sequence a_1^j of outcomes of $\mathcal{X}$, if some assumptions are made about the stochastic process which generated the sequence, the cross entropy $H_m(p_m, a_1^j)$ of model m with respect to event sequence a_1^j may be calculated as shown in Equation 6.6.[1]

$$\begin{aligned} H_m(p_m, a_1^j) &= -\frac{1}{j} \log_2 p_m(a_1^j) \\ &= -\frac{1}{j} \sum_{i=1}^{j} \log_2 p_m(a_i | a_1^{i-1}) \end{aligned} \tag{6.6}$$

Cross entropy approaches the true entropy of the sequence as the length of the sequence (j) increases.

Since $H_m(p_m, a_1^j)$ provides an estimate of the number of binary bits required on average to encode a symbol in a_1^j in the most efficient manner and there exist techniques, such as arithmetic coding (Witten *et al.*, 1987), which can produce near-optimal codes, cross entropy provides a direct performance metric in the realm of data compression. However, cross entropy has a wider use in the evaluation of statistical models. Since it provides a measure of how uncertain a model is, on average, when predicting a given sequence of events, it can be used to compare the performance of different models on some corpus of data. In statistical language modelling, cross entropy measures are commonly used:

> For a number of natural language processing tasks, such as speech recognition, machine translation, handwriting recognition, stenotype transcription and spelling correction, language models for which the cross entropy is lower lead directly to better performance.
>
> (Brown *et al.*, 1992, p. 39).

A related measure, *perplexity*, is also frequently used in statistical language modelling. The perplexity $Perplexity_m(p_m, a_1^j)$ of model m on sequence a_1^j is defined as:

$$Perplexity_m(p_m, a_1^j) = 2^{H_m(p_m, a_1^j)} \tag{6.7}$$

Perplexity provides a crude measure of the average size of the set of symbols

[1] In particular, it is standard to assume that the process is *stationary* and *ergodic* (Manning & Schütze, 1999). A stochastic process is stationary if the probability distribution governing the emission of symbols is stationary over time (*i.e.,* independent of the position in the sequence) and ergodic if sufficiently long sequences of events generated by it can be used to make inferences about its typical behaviour.

from which the next symbol is chosen – lower perplexities indicate better model performance.

6.2.3 The PPM Algorithm

6.2.3.1 Overview

Prediction by Partial Match (Cleary & Witten, 1984) is a data compression scheme of which the central component is an algorithm for performing backoff smoothing of n-gram distributions. Variants of the PPM scheme have consistently set the standard in lossless data compression since its original introduction (Bunton, 1997). Several of these variants will be described in terms of Equations 6.1 and 6.2 where the recursion is terminated with a model which returns a uniform distribution over $[\tau]$. This model is usually referred to as the *order minus-one* model and allows for the prediction of events which have yet to be encountered.

6.2.3.2 The Zero-frequency Problem and Escaping

In this section, the calculation of the probability estimates $\alpha()$ and $\gamma()$ in Equations 6.1 and 6.2 in PPM models is discussed. The problem is usually characterised by asking how to estimate the *escape probability* $\gamma(e_i|e_{(i-n)+1}^{i-1})$ which represents the amount of probability mass to assign to events which are novel in the current context $e_{(i-n)+1}^{i-1}$. The probability estimate $\alpha(e_i|e_{(i-n)+1}^{i-1})$ is then set such that the estimated distributions sum to unity. As noted by Witten & Bell (1991), there is no sound theoretical basis for choosing these *escape probabilities* in the absence of *a priori* knowledge about the data being modelled. As a result, although several schemes exist, their relative performance on any particular real-world task can only be determined experimentally. In the following discussion, $t(e_i^j)$ denotes the total number of *symbol types*, members of $[\tau]$, that have occurred with non-zero frequency in context e_i^j; and $t_k(e_i^j)$ denotes the total number of symbol types that have occurred exactly k times in context e_i^j.

Method A (Cleary & Witten, 1984) assigns a frequency count of one to symbols that are novel in the current context $e_{(i-n)+1}^{i-1}$ and adjusts $\alpha(e_i|e_{(i-n)+1}^{i-1})$ accordingly:

$$\gamma(e_i|e_{(i-n)+1}^{i-1}) = \frac{1}{\sum_{e\in[\tau]} c(e|e_{(i-n)+1}^{i-1}) + 1}$$
$$\alpha(e_i|e_{(i-n)+1}^{i-1}) = \frac{c(e_i|e_{(i-n)+1}^{i-1})}{\sum_{e\in[\tau]} c(e|e_{(i-n)+1}^{i-1}) + 1}$$

As the number of occurrences of the context increases, $\gamma(e_i|e_{(i-n)+1}^{i-1})$ decreases and $\alpha(e_i|e_{(i-n)+1}^{i-1})$ approaches the ML estimate.

Method B (Cleary & Witten, 1984) classifies a symbol occurring in a given context as novel unless it has already occurred *twice* in that context. This has the effect of filtering out anomalies and is achieved by subtracting one from the symbol counts when calculating $\alpha(e_i|e_{(i-n)+1}^{i-1})$. In addition, the appearance of the type count $t(e_{(i-n)+1}^{i-1})$ in the numerator of $\gamma(e_i|e_{(i-n)+1}^{i-1})$ has the effect that the escape probability increases as more types are observed.

$$\gamma(e_i|e_{(i-n)+1}^{i-1}) = \frac{t(e_{(i-n)+1}^{i-1})}{\sum_{e\in[\tau]} c(e|e_{(i-n)+1}^{i-1})}$$
$$\alpha(e_i|e_{(i-n)+1}^{i-1}) = \frac{c(e_i|e_{(i-n)+1}^{i-1}) - 1}{\sum_{e\in[\tau]} c(e|e_{(i-n)+1}^{i-1})}$$

Method C (Moffat, 1990) was designed to combine the more attractive elements of methods A and B. It is a modified version of method A in which the escape count increases as more types are observed (as in method B).

$$\gamma(e_i|e_{(i-n)+1}^{i-1}) = \frac{t(e_{(i-n)+1}^{i-1})}{\sum_{e\in[\tau]} c(e|e_{(i-n)+1}^{i-1}) + t(e_{(i-n)+1}^{i-1})}$$
$$\alpha(e_i|e_{(i-n)+1}^{i-1}) = \frac{c(e_i|e_{(i-n)+1}^{i-1})}{\sum_{e\in[\tau]} c(e|e_{(i-n)+1}^{i-1}) + t(e_{(i-n)+1}^{i-1})}$$

One particular smoothing technique called *Witten-Bell smoothing*, often used in statistical language modelling, is based on escape method C (Manning & Schütze, 1999).

Method D (Howard, 1993) modifies method B by subtracting 0.5 (instead of 1) from the symbol count when calculating $\alpha(e_i|e_{(i-n)+1}^{i-1})$.

$$\gamma(e_i|e_{(i-n)+1}^{i-1}) = \frac{\frac{1}{2}t(e_{(i-n)+1}^{i-1})}{\sum_{e\in[\tau]} c(e|e_{(i-n)+1}^{i-1})}$$
$$\alpha(e_i|e_{(i-n)+1}^{i-1}) = \frac{c(e_i|e_{(i-n)+1}^{i-1}) - \frac{1}{2}}{\sum_{e\in[\tau]} c(e|e_{(i-n)+1}^{i-1})}$$

Method AX (Moffat *et al.*, 1998) is motivated by the assumption that novel events occur according to a Poisson process model. On this basis, Witten & Bell (1991) have suggested method P which uses the following escape probability:

$$\gamma(e_i|e_{(i-n)+1}^{i-1}) = \frac{t_1(e_i|e_{(i-n)+1}^{i-1})}{\sum_{e\in[\tau]} c(e|e_{(i-n)+1}^{i-1})} - \frac{t_2(e_i|e_{(i-n)+1}^{i-1})}{(\sum_{e\in[\tau]} c(e|e_{(i-n)+1}^{i-1}))^2}\cdots$$

and method X which approximates method P by computing only the first term:

$$\gamma(e_i|e_{(i-n)+1}^{i-1}) = \frac{t_1(e_i|e_{(i-n)+1}^{i-1})}{\sum_{e\in[\tau]} c(e|e_{(i-n)+1}^{i-1})}$$

However, both of these methods suffer from the fact that when $t_1(e_i|e_{(i-n)+1}^{i-1}) = 0$ or $t_1(e_i|e_{(i-n)+1}^{i-1}) = \sum_{e\in[\tau]} c(e|e_{(i-n)+1}^{i-1})$, the escape probability will be zero (or less) or one respectively. One solution to this problem, suggested by Moffat *et al.* (1998) and dubbed method AX (for Approximate X), is to add one to the counts and use the singleton type count in method C.

$$\gamma(e_i|e_{(i-n)+1}^{i-1}) = \frac{t_1(e_{(i-n)+1}^{i-1}) + 1}{\sum_{e\in[\tau]} c(e|e_{(i-n)+1}^{i-1}) + t_1(e_{i-1}^{(i-n)+1}) + 1}$$
$$\alpha(e_i|e_{(i-n)+1}^{i-1}) = \frac{c(e_i|e_{(i-n)+1}^{i-1})}{\sum_{e\in[\tau]} c(e|e_{(i-n)+1}^{i-1}) + t_1(e_{(i-n)+1}^{i-1}) + 1}$$

These methods are based on similar principles to *Katz backoff* (Katz, 1987) one of the more popular smoothing techniques used in statistical language processing.

These various escape methods have been subjected to empirical evaluation in data compression experiments. In general, A and B tend to perform poorly (Bunton, 1997; Moffat *et al.*, 1994; Witten & Bell, 1991), while D tends to

slightly outperform C (Bunton, 1997; Moffat *et al.*, 1994) and methods based on P (*e.g.,* AX) tend to produce the best results (Moffat *et al.*, 1994; Teahan & Cleary, 1997; Witten & Bell, 1991).

6.2.3.3 Exclusion

Exclusion (Cleary & Witten, 1984) is a technique for improving the probabilities estimated by PPM based on the observation that events which are predicted at higher order contexts do not need to be included in the calculation of lower order estimates. Exclusion of events which have already been predicted in a higher level context will have no effect on the outcome (since they have already been predicted) and doing so reclaims a proportion of the overall probability mass that would otherwise be wasted. Unless explicitly stated otherwise, it is henceforth assumed that exclusion is enabled in all models discussed.

6.2.3.4 Interpolated Smoothing

The difference between backoff and interpolated smoothing was discussed in §6.2.1 where both kinds of smoothing were expressed within the same framework. While the original PPM algorithm uses a backoff strategy (called *blending*), Bunton (1996, 1997) has experimented with using interpolated smoothing within PPM. The approach is best described by rewriting Equation 6.2 such that:

$$\begin{aligned} \alpha(e_i|e_{(i-n)+1}^{i-1}) &= \lambda(e_{(i-n)+1}^{i-1}) \cdot \frac{count(e_i, e_{(i-n)+1}^{i-1})}{count(e_{(i-n)+1}^{i-1})} \\ \gamma(e_i|e_{(i-n)+1}^{i-1}) &= (1-\lambda)(e_{(i-n)+1}^{i-1}) \end{aligned}$$

where:

$$\begin{aligned} count(e_i, e_{(i-n)+1}^{i-1}) &= \begin{cases} c(e_i|e_{(i-n)+1}^{i-1}) + k & \text{if } c(e_i|e_{(i-n)+1}^{i-1}) > 0 \\ 0 & \text{otherwise} \end{cases} \\ count(e_{(i-n)+1}^{i-1}) &= \sum_{e \in [\tau]:\ e\ is\ not\ excluded} count(e, e_{(i-n)+1}^{i-1}) \end{aligned}$$

and k is the initial event frequency count and a global constant (ideally $k = 0$). The resulting smoothing mechanism is described in Equation 6.8 which computes the estimated probability of an n-gram consisting of a context s and a single event prediction e where $su(e_i^j) = e_{i+1}^j$.

$$p(e|s) = \begin{cases} \lambda(s) \cdot \frac{count(e,s)}{count(s)} + (1 - \lambda(s)) \cdot p(e|su(s)) & \text{if } su(s) \neq \varepsilon \\ \frac{1}{|[\tau]|+1-t(\varepsilon)} & \text{otherwise} \end{cases} \quad (6.8)$$

When using the interpolated smoothing described in Equation 6.8, it is difficult to ensure that the conditional probability distribution computed sums to unity. A simple, though computationally expensive, solution to this problem is to compute the entire distribution and then renormalise its component probabilities such that they do sum to unity.

As noted by Bunton (1996, ch. 6), methods A through D may be described using a single weighting function $\lambda : [\tau]^* \rightarrow [0, 1)$, defined as follows:

$$\lambda(e^{i-1}_{(i-n)+1}) = \frac{count(e^{i-1}_{(i-n)+1})}{count(e^{i-1}_{(i-n)+1}) + \frac{t(e^{i-1}_{(i-n)+1})}{d(e^{i-1}_{(i-n)+1})}}$$

if we allow the escape method to determine the values of k and a variable $d(e^{i-1}_{(i-n)+1})$ as follows:

$$\begin{aligned} \mathbf{A}: \; d(e^{i-1}_{(i-n)+1}) &= t(e^{i-1}_{(i-n)+1}), & k = 0; \\ \mathbf{B}: \; d(e^{i-1}_{(i-n)+1}) &= 1, & k = -1; \\ \mathbf{C}: \; d(e^{i-1}_{(i-n)+1}) &= 1, & k = 0; \\ \mathbf{D}: \; d(e^{i-1}_{(i-n)+1}) &= 2, & k = -\frac{1}{2}. \end{aligned}$$

Furthermore, method AX may be described within the same framework as follows:

$$\begin{aligned} \lambda(e^{i-1}_{(i-n)+1}) &= \frac{count(e^{i-1}_{(i-n)+1})}{count(e^{i-1}_{(i-n)+1}) + \frac{t_1(e^{i-1}_{(i-n)+1})}{d(e^{i-1}_{(i-n)+1})}} \\ d(e^{i-1}_{(i-n)+1}) &= 1 \\ k &= 0 \end{aligned}$$

Bunton (1996) observes that the key difference between escape methods A through D is the relative emphasis placed on lower and higher order distribu-

tions. More emphasis is placed on higher order distributions as both k and $d(e_{(i-n)+1}^{i-1})$ increase in numerical value. Thus, while method B places the lowest relative emphasis on higher order distributions, method A tends to place the greatest emphasis on higher order distributions (depending on the value of $d(e_{(i-n)+1}^{i-1}) = t(e_{(i-n)+1}^{i-1})$). Methods C, D and AX fall in between these extremes of emphasis and consistently outperform A and B in data compression experiments.

Blending drops a term of Equation 6.8 for events which are not novel by assuming that $p(e_i|e_{(i-n)+2}^{i-1}) = 0$. As discussed in in §6.2.1, this is true of backoff versions of interpolated smoothing methods in general. Bunton notes that, as a consequence, the estimates for novel events are slightly inflated while the estimates for events which are not novel are slightly deflated. Replacing blending with interpolated smoothing remedies this and yields significant and consistent improvements in compression performance (Bunton, 1996, 1997).

6.2.3.5 Update Exclusion

Update exclusion (Moffat, 1990) is a modified strategy for updating the n-gram counts in PPM models. When using the original PPM model with blending and exclusion, the probability of an event which is not novel in a given context, will be estimated in that context alone without blending the estimate with lower order estimates. Update exclusion refers to a counting strategy in which the event counts are only incremented if an event is not predicted in a higher order context. This has the effect that the counts more accurately reflect which events are likely to have been excluded in higher order contexts. The use of update excluded counts tends to improve the data compression performance of PPM models (Bell *et al.*, 1990; Bunton, 1997; Moffat, 1990).

6.2.3.6 Unbounded Length Contexts

One of the goals of *universal* modelling is to make minimal assumptions about the nature of the stochastic processes (or source) responsible for generating observed data. As discussed in §6.2.1, n-gram models make assumptions about a source to the effect that the probability of an event depends only on the previous $n-1$ events. Cleary & Teahan (1997) describe an extension to PPM, called PPM*, which eliminates the need to impose an arbitrary order bound. The policy used to select a maximum order context can be freely varied depending on the situation.

A context e_i^j is said to be *deterministic* when it makes exactly one prediction: $t(e_i^j) = 1$. Cleary & Teahan (1995) have found that for such contexts the

observed frequency of novel events is much lower than expected based on a uniform prior distribution. As a consequence, the entropy of the distributions estimated in deterministic contexts tend to be lower than in non-deterministic contexts. Since an event will have occurred at least as many times in the lowest order matching deterministic context as any of the other matching deterministic contexts, this context will yield the lowest-entropy probability distribution (Bunton, 1997). Cleary & Teahan (1997) exploit this in PPM* by selecting the shortest deterministic matching context if one exists or otherwise selecting the longest matching context. Unfortunately, the original PPM* implementation provided (at best) modest improvement in compression performance over the original order bounded PPM. When combined with interpolated smoothing and update exclusion, however, PPM* does outperform the corresponding order bounded PPM models in data compression experiments (Bunton, 1997). Furthermore, Bunton (1997) describes an information-theoretic state selection mechanism which yields additional improvements in the compression performance of PPM* models.

As noted by Bunton (1997), PPM*'s state selection mechanism interferes with the use of update excluded frequency counts since PPM* does not always estimate the probability distribution using the frequency data from the maximum order matching context. The solution is to use full counts to compute probabilities for the selected context and update excluded counts thereafter for the lower order contexts (see Bunton, 1996, 1997, for further details).

6.2.3.7 Implementation Issues

Since PPM* does not impose an order bound, all subsequences of the input sequence must be stored, which makes for increased demands on computational resources. Suffix-tree representations provide a space-efficient means of achieving this end (Bunton, 1996; Larsson, 1996). In the present research, PPM models have been implemented as suffix trees using the online construction algorithm described by Ukkonen (1995). The application of this algorithm to the construction of PPM models was first described by Larsson (1996) and the construction developed independently by Bunton (1996) is similar to the Ukkonen-Larsson algorithm in many respects. In addition to being online, these algorithms have linear time and space complexity and, as demonstrated by Bunton (1996), the resulting models have optimal space requirements (in contrast to the original PPM* implementation). Since the suffix trees developed in the present research are constructed from more than one sequence, they are in fact *generalised* suffix trees which require only minor modifications to Ukkonen's

suffix tree construction algorithm (Gusfield, 1997). The existence of path compressed nodes in suffix trees complicates the storage of frequency counts and their use in prediction. In the present research, these complications were addressed by following the strategies for initialising and incrementing the counts developed by Bunton (1996).

6.2.4 Long- and Short-term Models

In data compression, a model which is initially empty is constructed incrementally as more of the input data is seen. However, experiments with PPM using an initial model that has been derived from a training text demonstrate that pre-training the model, both with related and with unrelated texts, significantly improves compression performance (Teahan, 1998; Teahan & Cleary, 1996). A complementary approach is often used in the literature on statistical language modelling where improved performance is obtained by augmenting n-gram models derived from the entire training corpus with *cache* models which are constructed dynamically from a portion of the recently processed text (Kuhn & De Mori, 1990).

Conklin (1990) has employed similar ideas with music data by using both a *long-term model* (LTM) and a *short-term model* (STM). The LTM parameters are estimated on the entire training corpus and new data is added to the model after it is predicted on a composition-by-composition basis. The STM, on the other hand, is constructed online for each composition in the test set and is discarded after the relevant composition has been processed. The predictions of both models are combined to provide an overall probability estimate for the current event. The motivation for doing so is to take advantage of recently occurring n-grams whose structure and statistics may be specific to the individual composition being predicted.

Let τ_b be the basic event attribute currently under consideration and $[\tau_b] = \{t_1, t_2, \ldots, t_k\}$ its domain. In this chapter, $\tau_b = \texttt{cpitch}$, a basic attribute representing chromatic pitch (see §5.3). Let M be a set $\{ltm, stm\}$ containing the LTM and STM, and $p_m(t)$ be the probability assigned to symbol $t \in [\tau_b]$ by model $m \in M$. Perhaps the simplest method of combining distributions is to compute the arithmetic mean of the estimated probabilities for each symbol $t \in [\tau_b]$ such that:

$$p(t) = \frac{1}{|M|} \sum_{m \in M} p_m(t)$$

This combination technique may be improved by weighting the contributions made by each of the models such that:

$$p(t) = \frac{\sum_{m \in M} w_m p_m(t)}{\sum_{m \in M} w_m}$$

A method for calculating the weights, w_m, is described by Conklin (1990). It is based on the entropies of the distributions generated by the individual models such that greater entropy (and hence uncertainty) is associated with a lower weight. The weight of model m is calculated as shown in Equation 6.9.

$$w_m = H_{relative}(p_m)^{-b} \tag{6.9}$$

The *relative entropy* $H_{relative}(p_m)$ of a model is given by:

$$H_{relative}(p_m) \quad = \quad \begin{cases} H(p_m)/H_{max}(p_m) & \text{if } H_{max}([\tau_b]) > 0 \\ 1 & \text{otherwise} \end{cases}$$

where H and H_{max} are as defined in Equations 6.3 and 6.4 respectively. The bias $b \in \mathbb{Z}^+$ is a parameter giving an exponential bias towards models with lower relative entropy. Note that when $b = 0$, the weighted arithmetic scheme is equivalent to its unweighted counterpart since all models are assigned an equal weight of one. This weighting mechanism is described in more detail by Conklin (1990) who used the weighted arithmetic mean for combining both viewpoint predictions and the predictions of the long- and short-term models (see §7.3). Conklin & Witten (1995) used this method for combining viewpoint predictions only. In both cases, the combined use of long- and short-term models yields better prediction performance than either the LTM or STM used individually (Conklin, 1990).[2]

6.3 Experimental Methodology

6.3.1 Model Parameters

A PPM model has been implemented in Common Lisp such that each of the variant features described in §6.2.3 may be independently selected. The following shorthand will be used to describe the cross product of model parameters:

[2] Other combination techniques are discussed in §7.3.

Model type: indicated by 'LTM' and 'STM' for the long- and short-term models respectively while 'LTM+' indicates a long-term model in which new data is added to the LTM online as each new event is predicted;[3]

Escape method: indicated explicitly by 'A', 'B', 'C', 'D' or 'X' (the latter as a shorthand for method AX);

Order bound: indicated by an integer or '*' if unbounded;

Update exclusion: the use of update excluded counts is indicated by 'U' – update excluded counts are disabled by default;

Interpolated smoothing: PPM's blending is the default while the use of interpolated smoothing is indicated by an 'I'.

Thus, for example, a PPM long-term model with escape method C, unbounded order, update exclusion enabled and interpolated smoothing is denoted by 'LTMC*UI'. When it is clear which model is being referred to, we shall, for the sake of readability, drop the model type. When combined with a short-term model with the same parameters, the model would be denoted by 'LTMC*UI—STMC*UI' (for readability the two models are separated by a dash). It will be clear that the space of possible parameterisations of the model is very large indeed (even when the range of possible order bounds is limited). As a consequence of this large parameter space, the techniques have been applied incrementally, typically taking the best performing model in one experiment as the starting point for the next. Note that while the resulting models should reflect local optima in the parameter space, they are not guaranteed to be globally optimal.

Conklin & Witten (1995) used a PPM model to predict 100 chorale melodies harmonised by J. S. Bach (see §3.4). Note that this dataset is almost disjoint from Dataset 2 used in this research (see §4.3). The escape method used was B and both long- and short-term models were employed. The global order bounds of the LTM and STM were set at 3 and 2 respectively and the predictions combined using a Dempster-Shafer scheme (see §6.2.4 and §7.3). This model is described in the current scheme as LTMB3—STMB2. A multiple viewpoint system consisting of `cpitch` alone yielded a cross entropy of 2.05 bits per event. Using the weighted arithmetic combination method described in §6.2.4 for combining viewpoint predictions, Conklin & Witten (1995) were able to obtain cross entropy measures as low as 1.87 bits per event using more complex

[3] In the present research, new data is added to the LTM on an event-by-event basis rather than the composition-by-composition basis adopted by Conklin (1990).

multiple viewpoint systems. On the basis of empirical results and theoretical considerations discussed in §6.2, the following predictions are made regarding model performance: the combined use of an STM and LTM will yield improved performance over either model used in isolation; the use of PPM* with interpolated smoothing will yield performance improvements over order-bounded PPM models using blending; using update excluded counts will improve performance over the standard counting strategy; and finally, escape methods C, D or AX will result in performance improvements over methods A and B.

6.3.2 Performance Evaluation

Many methods have been used to evaluate the performance of statistical models of music, some of which have been described in §3.4: the analysis-by-synthesis method used by Hall & Smith (1996) and Triviño-Rodriguez & Morales-Bueno (2001); comparison of human and machine prediction performance (Witten *et al.*, 1994); single-sample Bayesian methods such as Minimum Description Length (Conklin, 1990); and the resampling approach using entropy as a measure of performance as used by Conklin & Witten (1995) and Reis (1999). The latter approach is followed here for two reasons: first, entropy has an unambiguous interpretation in terms of model uncertainty on unseen data (see §6.2.2); and second, entropy bears a direct relationship to performance in data compression, correlates with the performance of n-gram models on a range of practical natural language tasks (Brown *et al.*, 1992) and is widely used in both these fields (see §6.2.2). These factors support its use in an application-independent evaluation such as this.

Conklin & Witten (1995) used a *split-sample* (or *held-out*) experimental paradigm in which the data is divided randomly into two disjoint sets, a training set and a test set. The n-gram parameters are then estimated on the training set and the cross entropy of the test set given the resulting model is computed using Equation 6.6 (see §6.2.2). Conklin & Witten divided their set of 100 chorale melodies into a training set of 95 melodies and a test set of 5 melodies. Although commonly used, split-sample validation suffers from two major disadvantages: first, it reduces the amount of data available for both training and testing; and second, with small datasets it provides a biased estimate of the true entropy of the corpus. A simple way of addressing these limitations is to use *k-fold cross-validation* (Dietterich, 1998; Kohavi, 1995a; Mitchell, 1997) in which the data is divided into k disjoint subsets of approximately equal size. The model is trained k times, each time leaving out a different subset to be used for testing and an average of the k cross entropy values thus obtained is

	Training set		Test set	
ID	Mean Compositions	Mean Events	Mean Compositions	Mean Events
1	136.8	7697.7	15.2	855.3
2	166.5	8304.3	18.5	922.7
3	81.9	4046.4	9.1	449.6
4	107.1	2421.9	11.9	269.1
5	83.7	4127.4	9.3	458.6
6	93.6	4775.4	10.4	530.6
7	191.7	7553.7	21.3	839.3
8	213.3	9950.4	23.7	1105.6

Table 6.1: The average sizes of the resampling sets used for each dataset.

then computed.

The data used in the experiments consisted of Datasets 1–8 (see Chapter 4). These experiments were carried out with a single viewpoint system consisting of a viewpoint for the basic type `cpitch` alone (the extension of the model to multiple viewpoint systems is presented in Chapter 7). Since the datasets used are quite small and initial experiments demonstrated a fairly large variance in the entropies computed from different validation sets, 10-fold cross-validation over each dataset was used in all experiments. The value of $k = 10$ is chosen as a commonly used compromise between the bias associated with low values of k and the high variance associated with high values of k (Kohavi, 1995b, ch. 3). The average sizes of the training and test sets for each dataset are shown in Table 6.1. Since the 10-fold partitioning of each dataset was achieved randomly, there is no reason to *expect* that the results will be different with alternative partitions. In machine learning research, differences in model performance as assessed by resampling techniques, such as cross-validation, are often analysed for significance using statistical tests such as the t test (Dietterich, 1998; Mitchell, 1997). This approach is followed in §6.4.4 where the overall performance improvements obtained using the methods described in §6.2.3 are examined in relation to an emulation of the model developed by Conklin & Witten (1995).

6.4 Results

6.4.1 Global Order Bound and Escape Method

The first experiments address the question of how the performance of PPM models is affected by changes in the global order bound. Both the LTM and STM

have been examined independently with all five escape methods using global order bounds ranging from zero to 14. The results for the LTM and STM are shown in Figures 6.1 and 6.2 respectively. The general *U*-shape of the curves is quite typical; while increasing the global order bound provides the model with more specific contextual information with which to make its predictions, the higher order contexts are also more likely to fail to produce a prediction. Therefore, the model will escape down to lower order models more frequently, thereby wasting more of the probability mass available on apportioning escape probabilities. As the global order bound is increased beyond a certain point this negative influence tends to dominate and performance decreases (Teahan, 1998).

Note, however, the relatively shallow degradation of performance of the STM (as compared with the LTM) as the global order bound is increased beyond its optimal value. It seems likely that due to the short length of most of the compositions in the datasets (see Chapter 4), the models rarely encounter matching contexts longer than about five events and, as a consequence, increasing the global order bound beyond this value has little effect on model performance. Figures 6.1 and 6.2 also reveal, for the LTM and STM respectively, that escape methods A and B perform relatively poorly and escape method C outperforms the others. It seems likely that the especially poor performance of method B with the STM arises from the small amount of data available for training combined with the fact that method B classifies a symbol as novel unless it has occurred twice in a given context (see §6.2.3.2). As a consequence of the poor performance of methods A and B, which corroborates findings in data compression experiments (Bunton, 1997), these two escape methods are not considered further in the present research.

The results also show some interesting trends regarding global order bound. As discussed in §6.2.3.6, the optimal global order bound to use is highly dependent on the amount and character of the data being used (Bunton, 1997). Using the present corpus of melodies Figures 6.1 and 6.2 demonstrate, respectively, that the LTM operates best with a global order bound of two, regardless of the escape method used, while the STM performs best with a global order bound of five with escape methods D and C and a global order bound of four with escape method AX.

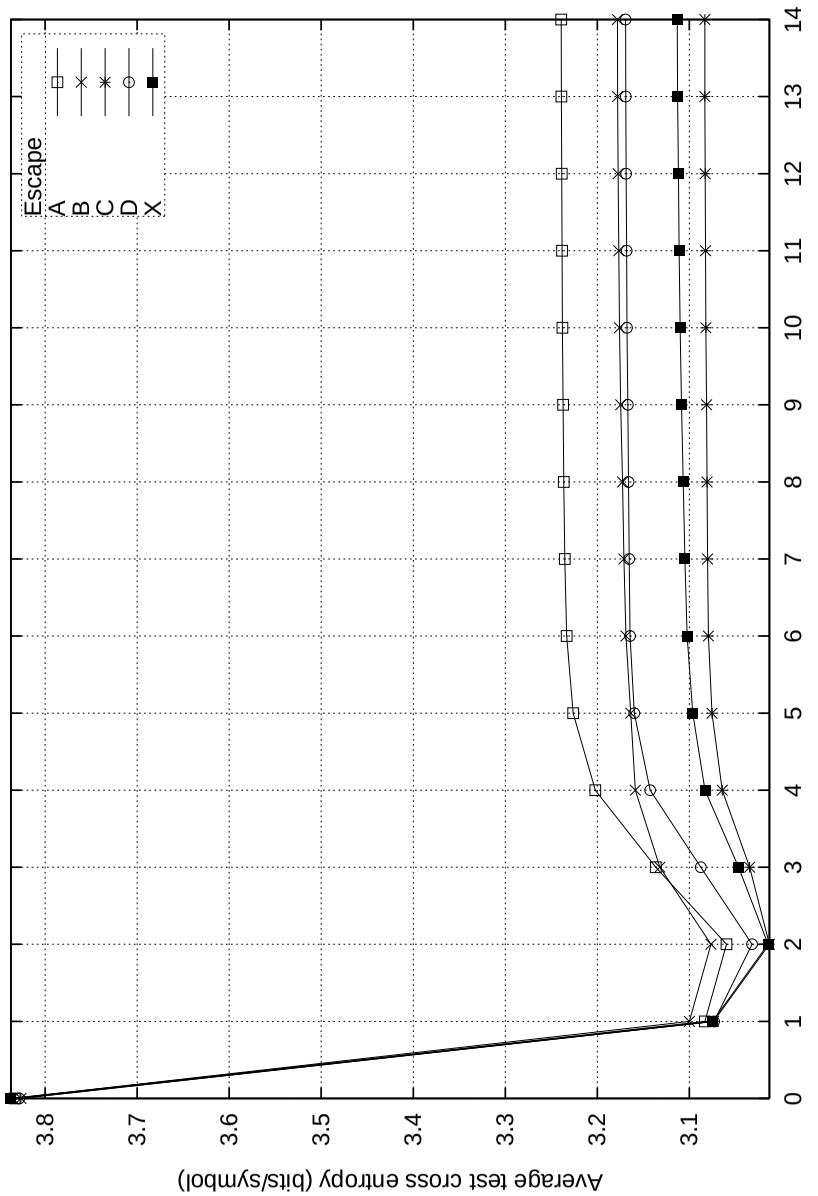


Figure 6.1: The performance of the LTM with varying escape method and global order bound.

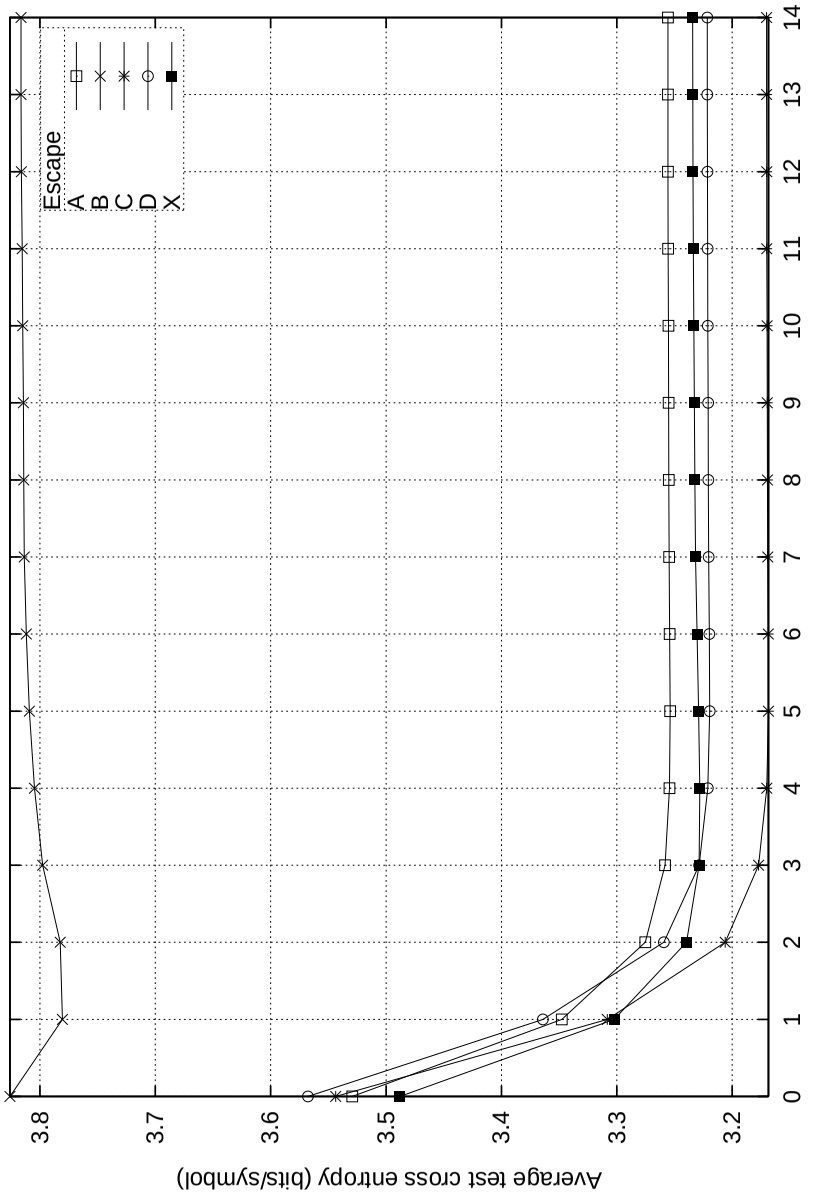


Figure 6.2: The performance of the STM with varying escape method and global order bound.

Dataset	C2	C2U	C2I	C2UI	D2	D2U	D2I	D2UI	X2	X2U	X2I	X2UI
1	2.933	2.959	2.904	3.127	2.935	2.951	**2.885**	2.967	2.913	2.928	2.887	2.908
2	2.585	2.595	2.563	2.748	2.577	2.581	2.547	2.608	2.557	2.562	**2.544**	2.554
3	3.216	3.204	**3.110**	3.417	3.252	3.208	3.142	3.220	3.207	3.161	3.166	3.129
4	2.882	2.890	2.804	3.179	2.892	2.881	**2.791**	2.954	2.880	2.870	2.824	2.829
5	3.276	3.248	**3.192**	3.483	3.315	3.250	3.220	3.278	3.312	3.231	3.277	3.201
6	3.470	3.480	**3.385**	3.708	3.526	3.485	3.431	3.509	3.518	3.455	3.485	3.429
7	2.620	2.665	2.613	2.897	2.622	2.654	2.599	2.731	2.608	2.642	**2.596**	2.633
8	3.123	3.157	**3.083**	3.423	3.137	3.145	3.094	3.203	3.121	3.123	3.111	3.111
Average	3.013	3.025	**2.957**	3.248	3.032	3.019	2.964	3.059	3.014	2.997	2.986	2.974

Table 6.2: Performance of the LTM with a global order bound of two.

Dataset	C5	C5U	C5I	C5UI	D5	D5U	D5I	D5UI	X4	X4U	X4I	X4UI
1	3.017	3.046	2.988	2.993	3.068	3.048	3.049	2.995	3.081	3.070	3.029	**2.983**
2	3.170	3.209	3.138	3.149	3.218	3.214	3.194	3.153	3.198	3.204	3.162	**3.121**
3	3.120	3.141	3.106	**3.104**	3.175	3.140	3.171	3.107	3.197	3.178	3.156	3.106
4	3.463	3.488	3.466	3.463	3.498	3.491	3.516	3.470	3.440	3.467	3.432	**3.411**
5	3.146	3.178	**3.134**	3.142	3.196	3.176	3.194	3.147	3.214	3.207	3.175	3.139
6	3.264	3.281	3.255	**3.252**	3.316	3.280	3.317	3.257	3.343	3.317	3.303	3.255
7	2.735	2.759	**2.701**	2.706	2.780	2.755	2.755	2.704	2.841	2.856	2.755	2.742
8	3.434	3.437	3.426	3.406	3.504	3.446	3.504	3.417	3.511	3.466	3.485	**3.402**
Average	3.169	3.192	3.152	3.152	3.220	3.194	3.213	3.156	3.228	3.220	3.187	**3.145**

Table 6.3: Performance of the STM with a global order bound of five (escape methods C and D) or four (escape method AX).

6.4.2 Interpolated Smoothing and Update Exclusion

The objective in the next experiments was to investigate the effects of using update excluded counts and interpolated smoothing on the performance of PPM models with optimal global order bounds as determined in the previous experiment. In these experiments, the STM and LTM were examined with escape methods C, D and AX using global order bounds of two for the LTM and five (escape methods C and D) or four (escape method AX) for the STM. The use of update excluded counts and interpolated smoothing were applied to these models both individually and in combination. The results for the LTM and STM are shown in Tables 6.2 and 6.3 respectively.

Consider first the results for the LTM shown in Table 6.2. Perhaps the most striking result is that interpolated smoothing applied in isolation improves performance for all datasets and escape methods. The best performing models on any given dataset use interpolated smoothing in isolation and, as in the previous experiment, escape method C tends on average to outperform methods D and AX. The results for update exclusion are, in general, more equivocal. Using update exclusion in isolation improves average model performance for escape methods D and AX but not for C (although the margin is small and performance is improved for Datasets 2 and 4). The combination of update exclusion and interpolated smoothing tends to impair performance, compared with the performance of models using either technique in isolation, for escape methods C and D; the slight average performance improvement with escape method AX derives from the improved performance on Datasets 2, 4 and 5.

The results for the STM, shown in Table 6.3, demonstrate that interpolated smoothing applied in isolation tends to improve performance though with less consistency across datasets and escape methods than it does with the LTM. By contrast, update exclusion (applied in isolation) improves average performance when used with escape methods D and AX but impairs performance with escape method C. Even more striking is the finding that the best average performance for each of the three escape methods is obtained using a combination of interpolated smoothing and update exclusion. However, the improvement over models using interpolated smoothing in isolation is much more pronounced for escape methods D and AX than for C where improvement is obtained for Datasets 2, 3, 5 and 7 only. The model with best average performance uses escape method AX with update exclusion and interpolated smoothing.

Dataset	C*	C*U	C*I	C*UI	D*	D*U	D*I	D*UI	X*	X*U	X*I	X*UI
1	3.094	3.236	**2.861**	3.234	3.180	3.247	2.930	3.098	3.072	3.153	2.933	2.993
2	2.669	2.843	**2.444**	2.869	2.708	2.839	2.473	2.724	2.648	2.812	2.477	2.651
3	3.336	3.407	**3.115**	3.470	3.454	3.424	3.230	3.308	3.320	3.315	3.230	3.166
4	2.937	3.032	**2.721**	3.188	3.004	3.040	2.761	2.998	2.965	3.028	2.809	2.862
5	3.176	3.199	**3.010**	3.316	3.293	3.205	3.119	3.147	3.263	3.176	3.187	3.056
6	3.515	3.550	**3.340**	3.645	3.665	3.562	3.486	3.488	3.606	3.482	3.542	3.370
7	2.604	2.779	**2.428**	2.926	2.681	2.780	2.468	2.739	2.614	2.748	2.480	2.593
8	3.318	3.449	**3.105**	3.556	3.395	3.434	3.188	3.347	3.298	3.348	3.189	3.205
Average	3.081	3.187	**2.878**	3.275	3.172	3.191	2.957	3.106	3.098	3.133	2.981	2.987

Table 6.4: Performance of the LTM with unbounded order.

Dataset	C*	C*U	C*I	C*UI	D*	D*U	D*I	D*UI	X*	X*U	X*I	X*UI
1	3.008	3.046	2.983	2.991	3.060	3.055	3.045	3.000	3.063	3.060	3.020	**2.977**
2	3.170	3.211	3.139	3.150	3.223	3.226	3.201	3.161	3.191	3.194	3.162	**3.117**
3	3.105	3.135	3.097	3.098	3.161	3.144	3.162	3.109	3.168	3.157	3.140	**3.090**
4	3.459	3.491	3.463	3.465	3.495	3.500	3.514	3.477	3.430	3.465	3.427	**3.411**
5	3.136	3.180	**3.126**	3.144	3.186	3.190	3.188	3.158	3.194	3.203	3.165	3.137
6	3.254	3.279	3.248	3.249	3.306	3.286	3.311	3.261	3.317	3.301	3.289	**3.244**
7	2.721	2.753	**2.693**	2.701	2.767	2.759	2.748	2.707	2.814	2.837	2.742	2.731
8	3.432	3.446	3.426	3.414	3.506	3.469	3.508	3.437	3.501	3.467	3.482	**3.406**
Average	3.161	3.192	3.147	3.152	3.213	3.203	3.210	3.164	3.210	3.211	3.179	**3.139**

Table 6.5: Performance of the STM with unbounded order.

6.4.3 Comparing PPM and PPM* Models

The objective in the next set of experiments was to investigate the effect of using update excluded counts and interpolated smoothing with (unbounded order) PPM* models with a view to comparing the unbounded models with their order-bounded counterparts. As in the previous experiments, the STM and LTM were tested with escape methods C, D and AX and were examined with update excluded counts and interpolated smoothing enabled both individually and in combination. The results for the LTM and STM are shown in Tables 6.4 and 6.5 respectively and exhibit broadly similar patterns to the corresponding order bounded results shown in Tables 6.2 and 6.3.

The results for the LTM shown in Table 6.4 demonstrate that, as in the order bounded experiment, interpolated smoothing (applied in isolation) universally improves performance. The use of update exclusion (applied in isolation) tends to impair performance, the only exceptions being when it was used in combination with escape methods D and AX on Datasets 2, 4 and 5. In combination with interpolated smoothing, update exclusion also tends to impair performance, the only exceptions being when it was used in combination with escape method AX on Datasets 2, 4 and 5. The trend for escape method C to outperform the other methods was stronger here than in the order bounded experiment and the best performing model on all datasets used interpolated smoothing and escape method C. Although the use of unbounded orders fails to consistently improve performance when the default blending scheme is used, the combination with interpolated smoothing does lead to consistent performance improvements over the corresponding order bounded models.

The results for the STM shown in Table 6.5 demonstrate that, as in the case of the order bounded STM results, interpolated smoothing applied in isolation tends to improve performance. The effect of update exclusion, both in isolation and in combination with interpolated smoothing, tends to be highly dependent both on the dataset and the escape method used. As in the order bounded experiment, escape methods D and AX tend to combine more fruitfully with update exclusion than method C. The models with best average performance for the former escape methods are obtained with a combination of update exclusion and interpolated smoothing. As in the order bounded experiment, the model with best average performance uses escape method AX with update exclusion and interpolated smoothing and this model outperforms its order-bounded counterpart.

	STMC*I	LTMC*I	LTM+C*I	LTM+C*I—STMC*I								
Dataset				b=0	b=1	b=2	b=3	b=4	b=5	b=6	b=16	b=32
1	2.983	2.861	2.655	2.495	2.475	**2.468**	2.469	2.474	2.482	2.491	2.564	2.608
2	3.139	2.444	2.375	2.396	2.363	2.347	**2.342**	**2.342**	2.346	2.352	2.412	2.455
3	3.097	3.115	2.712	2.554	2.541	**2.540**	2.548	2.559	2.571	2.584	2.677	2.730
4	3.463	2.721	2.602	2.619	2.597	**2.588**	2.589	2.595	2.604	2.614	2.714	2.791
5	3.126	3.010	2.621	2.484	2.461	**2.454**	2.457	2.465	2.474	2.485	2.560	2.610
6	3.248	3.340	2.833	2.659	**2.649**	2.651	2.661	2.675	2.690	2.706	2.816	2.880
7	2.693	2.428	2.237	2.153	2.120	2.106	**2.102**	2.104	2.109	2.116	2.176	2.212
8	3.426	3.105	2.881	2.694	**2.680**	2.681	2.691	2.705	2.720	2.735	2.841	2.902
Average	3.147	2.878	2.614	2.507	2.486	**2.479**	2.482	2.490	2.500	2.510	2.595	2.648

Table 6.6: Performance of the best performing long-term, short-term and combined models with variable bias.

6.4.4 Combining the Long- and Short-term Models

The objective of the next experiment was to examine the combined performance of the LTM and STM whose predictions were combined as described in §6.2.4. In general, the approach followed in these experiments has been to select the best performing models at any given stage for further experimentation. Accordingly, the LTMC*I model was chosen for use in these experiments (see §6.4.3). However, although the STMX*UI model was found to perform optimally in isolation (see §6.4.3), in a preliminary series of informal pilot experiments it was found that an STMC*I model yielded slightly better performance than a STMX*UI model in combination with the LTMC*I model. This finding in combination with the principle of Ockham's razor (the LTM and STM both use the same escape method) led us to select an STMC*I model over an STMX*UI model for use in these experiments.[4]

The results of this experiment are shown in Table 6.6. The first two columns respectively show the performance of the STMC*I and LTMC*I models used in isolation. The third column demonstrates the improved performance afforded by an LTM+C*I model in which events are added online to the LTM as they are predicted (see §6.2.4). The remainder of Table 6.6 shows the results obtained by combining the STMC*I model with the LTM+C*I model with a range of different values for the weighting bias b. As can be seen, a combined LTM—STM model is capable of outperforming both of its constituent models. The results also demonstrate that optimal average performance is achieved with the bias set to two.

[4]However, in the context of the multiple viewpoint system presented in Chapter 7 the STMX*UI model was found to improve performance over the STMC*I model.

Dataset	LTM+B3—STMB2	LTM+C3—STMC2	LTM+C*—STMC*	LTM+C*I—STMC*I
1	2.905	2.613	2.562	2.468
2	2.676	2.488	2.460	2.347
3	2.997	2.689	2.616	2.540
4	2.934	2.698	2.665	2.588
5	2.974	2.640	2.495	2.454
6	3.233	2.819	2.698	2.651
7	2.555	2.270	2.158	2.106
8	3.111	2.796	2.793	2.681
Average	2.923	2.627	2.556	2.479

Table 6.7: Performance improvements to an emulation of the model used by Conklin & Witten (1995).

6.4.5 Overall Performance Improvements

To illustrate more clearly the performance improvements obtained with the PPM variants discussed in this chapter, a final experiment was conducted in which escape method C, unbounded orders and interpolated smoothing were successively applied to an emulation of the model used by Conklin & Witten (1995) which is described in this framework as LTM+B3—STMB2 (see §6.3).[5] The results are shown in Table 6.7. Paired t tests confirmed the significance of the improvements afforded by incrementally applying escape method C [$t(79) = 31.128, p < 0.001$], unbounded orders [$t(79) = 9.018, p < 0.001$] and interpolated smoothing [$t(79) = 18.281, p < 0.001$]. The tests were performed over all 10 resampling sets of each dataset ($n = 80$) although, for reasons of space, Table 6.7 contains just the aggregate means for each of the eight datasets. The combined effect of the techniques applied in the LTMC*I—STMC*I model (shown in the final column of Table 6.7) is a 15% improvement in average model performance over the LTMB+3—STMB2 model used by Conklin & Witten (shown in in the first data column of Table 6.7).

[5] At the time of writing, there was insufficient information to enable a precise replication of the experiments described by Conklin & Witten (1995). Any discrepancy between the results reported here for Dataset 2 and those of Conklin & Witten may be attributed to several factors: first, the use by Conklin & Witten of a smaller, almost disjoint set of chorale melodies; second, the smaller pitch alphabet derived from this dataset; third, the use here of ten-fold cross-validation with an average of 18.5 compositions in the test set compared with the split sample paradigm employed by Conklin & Witten with a training set of 95 and test set of 5 compositions; and finally, the use of a Dempster-Shafer scheme by Conklin & Witten (see §7.3) for combining the predictions of the LTM and STM as compared with the weighted average employed here.

6.5 Discussion and Conclusions

Before discussing the results presented in §6.4, some words on the methodology employed are in order. The goal was to conduct an empirical test of the hypothesis that a number of techniques improve the prediction performance of PPM models on monophonic music data. This task has been approached by using cross entropy of the models as the performance metric and applying ten-fold cross validatory resampling on eight monophonic datasets. Since these experiments were concerned with optimising average performance over all eight datasets, the best performing models selected in some experiments (*e.g.,* the global order bound experiments described in §6.4.1) will not necessarily correspond to the best performing models on any single dataset. However, these best performing models inspire increased confidence that the model will perform well on a new dataset without requiring further empirical investigation of that dataset: *i.e.,* less information about the dataset is needed in advance to be confident of improved performance on that dataset.

The variant techniques have been applied incrementally, typically by taking the best performing model in a given experiment as the starting point for the next experiment. For example, in §6.4.4, the LTM and STM which yielded best performance independently were selected as the models to combine. Although there is no guarantee that the resulting model reflects the global optimum in the space of possible LTM and STM parameterisations, the objective was to demonstrate that some variant techniques can improve the performance of PPM models and consequently, the relative performance of the PPM variants is of more interest than their absolute performance. In this regard, it has been demonstrated that the combined use of three variant techniques affords significant and consistent performance improvements of 15% on average over the model used by Conklin & Witten (1995). The implications of the experimental results are now discussed in more detail for each of the variant techniques in turn.

Escape Method As noted in §6.2.3.2, there is no principled means of selecting the escape method (the probability to assign to events which have never arisen in a given context before) in the absence of *a priori* knowledge about the data. In the experiments reported here, escape methods A and B were consistently outperformed by C, D and AX and C fairly consistently outperformed both D and AX (although method AX performed well with the short-term model). These results are broadly in agreement with those obtained in data compression experiments (Bunton, 1996; Moffat *et al.*, 1994; Witten & Bell, 1991). Escape

method C is the most commonly used method when Witten-Bell smoothing is used in statistical language modelling (Manning & Schütze, 1999).

Interpolated Smoothing The use of interpolated smoothing consistently improves model performance (by comparison with PPM's default blending strategy) regardless of the dataset and combination with other variant techniques. This is consistent with results obtained in experiments in data compression (Bunton, 1997) and on natural language corpora (Chen & Goodman, 1999). The reason appears to derive from the fact that backoff smoothing (of which blending is an example) consistently underestimates the probabilities of non-novel events (Bunton, 1997) for which the low order distributions provide valuable information. For natural language corpora, this effect is particularly strong for n-grams with low frequency counts (Chen & Goodman, 1999).

Update Exclusion While update exclusion generally improves the performance of PPM models in data compression experiments (Bunton, 1997; Moffat, 1990), the results in these experiments were more equivocal. In general, the effects of update exclusion appeared to be highly sensitive to factors such as the dataset, escape method and model type (LTM or STM). In particular, escape methods AX, D and C respectively benefited less from the use of update excluded counts. Furthermore, the LTM appeared to benefit rather less from update exclusion than did the STM. Finally, when update exclusion did improve average performance, it tended to be the result of improvements on a restricted set of datasets. These findings are not entirely without precedent. The results presented by Bunton (1997) demonstrate that, although it improves average compression performance, update exclusion impairs performance for some of the test files and that escape method C benefits slightly less from the use of update excluded counts than method D.

Unbounded Orders The use of unbounded orders, as described in §6.2.3.6, failed to yield consistent improvements in performance for both the LTM and STM except when used in combination with interpolated smoothing. This combination of unbounded orders and interpolated smoothing, however, consistently improves the performance of the best performing order bounded models with interpolated smoothing. These results are consistent with those obtained in data compression experiments (Bunton, 1997) and this is likely to be due to the fact that the optimal order bound varies between datasets. As noted by Bunton (1997, p. 90), order bound experiments "provide more information

about the nature of the test data, rather than the universality of the tested algorithms." The advantage of PPM* is that it requires fewer assumptions to be made about the character of the data used.

Combined LTM and STM As expected on the basis of previous research (Conklin, 1990; Kuhn & De Mori, 1990; Teahan, 1998), combining the predictions of the LTM and STM improves model performance by comparison with that of either model used independently. Curiously, Conklin (1990) found that performance continued improving when the bias b was set to values as high as 128 and greater. In the experiments reported here, the optimal bias setting ranged from one to four depending on the dataset. Further experiments with the bias set to values as high as 32 only yielded further reduction in performance.

6.6 Summary

The research goal in this chapter was to evaluate, in an application independent manner, the performance improvements resulting from the application of a number of variant techniques to a class of n-gram models. In §6.2.1, n-gram modelling was introduced while in §6.2.2, the information-theoretic performance measures that have been used were described. Particular attention was given to PPM models in §6.2.3, where a number of techniques that have been used to improve the performance of PPM models were described in detail. These techniques include a range of different escape methods (§6.2.3.2), the use of update excluded counts (§6.2.3.5), interpolated smoothing (§6.2.3.4), unbounded orders (§6.2.3.6) and combining the predictions of a LTM and STM (§6.2.4). In a series of experiments, these techniques were applied incrementally to eight melodic datasets using cross entropy computed by 10-fold cross-validation on each dataset as the performance metric (see §6.3). The results reported in §6.4 demonstrate the consistent and significant performance improvements afforded by the use of escape method C (although method AX also performed well with the short-term model), unbounded orders, interpolated smoothing and combining long- and short-term models. Finally, in §6.5 the results were discussed in the context of previous research on the statistical modelling of music and in the fields of data compression and statistical language modelling.

CHAPTER 7

COMBINING PREDICTIVE MODELS OF MELODIC MUSIC

7.1 Overview

As described in Chapter 5, a multiple viewpoint representation scheme has been developed in the present research to address the need to flexibly represent many diverse attributes of the musical surface. In this chapter, the statistical modelling techniques presented in Chapter 6 are applied within the multiple viewpoint framework presented in §5.4. Multiple viewpoint modelling strategies take advantage of such a representational framework by deriving individual expert models for any given representational viewpoint and then combining the results obtained from each model (Conklin & Witten, 1995). The specific objective in this chapter is to evaluate methods for combining the predictions of different models in a multiple viewpoint system. To this end, the performance of the combination technique based on a weighted arithmetic mean described in §6.2.4 is compared with that of a new technique based on a weighted geometric mean. A second goal is to examine in greater detail the potential for multiple viewpoint systems to reduce model uncertainty in music prediction. A feature selection algorithm is used to derive a set of viewpoints selected from those described in Table 5.2 which optimises model uncertainty over a given corpus.

Multiple viewpoint systems are a specific instance of a more general class of strategies in machine learning collectively known as *ensemble learning methods*. As noted by Dietterich (2000), ensemble methods can improve the performance of machine learning algorithms for three fundamental reasons. The first is sta-

tistical: with small amounts of training data it is often hard to obtain reliable performance measures for a single model. By combining a number of well-performing models, it is possible to reduce the risk of inadvertently selecting models whose performance does not generalise well to new examples. The second reason is computational: for learning algorithms which employ local search, combining models which search locally from different starting points in the hypothesis space can yield better performance than any of the individual models. The final reason is representational: the combination of hypotheses drawn from a given space may expand the space of representable functions. The development of multiple viewpoint systems was motivated largely by representational concerns arising specifically in the context of computer modelling of music (Conklin & Witten, 1995). Although ensemble methods have typically been applied in classification problems, as opposed to the prediction problems studied here, research on ensemble methods in classification tasks will be drawn on as required.

The chapter is structured as follows. In §7.2, the development of statistical models within the multiple viewpoint framework is reviewed, while in §7.3, a new technique for combining viewpoint predictions is introduced. The experimental procedure used to evaluate this technique is described in §7.4 which also contains a description of the feature selection algorithm used to develop a multiple viewpoint system with reduced model uncertainty. Finally, the results of the experiments are presented and discussed in §7.5.

7.2 Background

7.2.1 Multiple Viewpoint Modelling of Music

For our purposes in this chapter, a statistical model associated with a viewpoint τ is a function m_τ which accepts a sequence of events in τ^* and which returns a distribution over $[\tau]$ reflecting the estimated conditional probabilities of the identity of the next viewpoint element in the sequence (see Chapter 6 for a detailed description of the model). A predictive system operating on a multiple viewpoint representation language consists of a number of models $m_{\tau_1}, \ldots, m_{\tau_n}$ corresponding to the collection of viewpoints $\tau_1, \ldots, \tau_n$ in the multiple viewpoint system. As described in §6.2.4, two models are actually employed for each viewpoint: a *long-term model* (LTM) and a *short-term model* (STM). As demonstrated in Chapter 6 the combined use of long- and short-term models yields better performance than that of either model used independently. The predictions of both long- and short-term models must be combined to produce

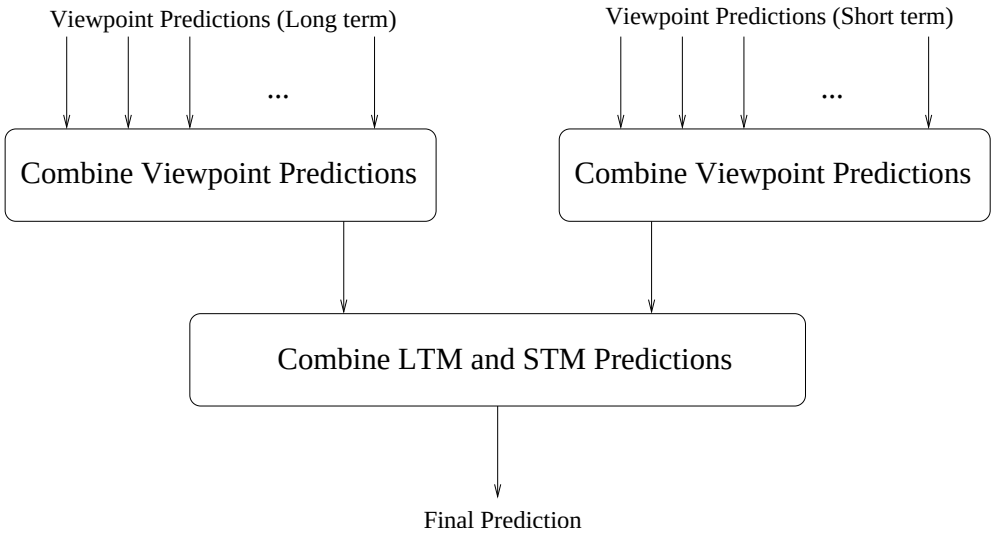


Figure 7.1: The architecture of a multiple viewpoint system (adapted from Conklin & Witten, 1995).

a final prediction (see §6.2.4). A number of general architectures can be envisaged to achieve this combination:

1. combine the STM and LTM predictions for each viewpoint individually and then combine the resulting viewpoint predictions;

2. combine the viewpoint predictions separately for the long- and short-term models and then combine the resulting LTM and STM predictions;

3. combine all long- and short-term viewpoint predictions in a single step.

The present research follows the practice of Conklin & Witten (1995) in choosing the second of these three architectures (see Figure 7.1). Two additional issues arise from the fact that the models accept sequences in $[\tau]^*$ (the set of viewpoint sequences over which the model was trained) rather than ξ^* (the set of basic event sequences) and return distributions over $[\tau]$ rather than ξ: first, the corpus of event sequences in ξ^* must be preprocessed into sequences in $[\tau]^*$ which are used to train the models; and second, the resulting distribution over $[\tau]$ must be postprocessed into a distribution over ξ so it may be combined with distributions generated by other models. These issues are discussed in §7.2.2 and §7.2.3 respectively.

7.2.2 Preprocessing the Event Sequences

Following Conklin & Witten (1995), sequences in ξ^* are converted to sequences in $[\tau]^*$ recursively using the function $\Phi_\tau : \xi^* \rightarrow [\tau]^*$ such that:

$$\Phi_\tau(e_1^j) = \begin{cases} \varepsilon & \text{if } e_1^j = \varepsilon \\ \Phi_\tau(e_1^{j-1}) & \text{if } \Psi_\tau(e_1^j) = \perp \\ \Phi_\tau(e_1^{j-1})\Psi_\tau(e_j) & \text{otherwise} \end{cases}$$

where Ψ_τ is the viewpoint function for attribute type τ which accepts a basic event sequence and returns an element of $[\tau]$. Since $\Psi_\tau(e_1^j) = \perp \Rightarrow \Phi_\tau(e_1^j) = \Phi_\tau(e_1^{j-1})$, it is necessary to check that $\Psi_\tau(e_1^j)$ is defined in order to prevent any sequence in $[\tau]^*$ being added to the model more than once (Conklin & Witten, 1995).

7.2.3 Completion of a Multiple Viewpoint System

A model m_τ returns a distribution over $[\tau]$ but, in order to combine the distributions generated by the models for different viewpoints, this must be converted into a distribution over the basic event space ξ. Any viewpoint which is undefined at the current location in the melody offers no information on which to make a prediction and is therefore eliminated from the combination process. In the interests of efficiency, prediction is elicited in stages, one for each basic type of interest (Conklin, 1990). Only those viewpoints which contain in their type set the basic type, τ_b, currently under consideration are activated at each stage. Linked or threaded viewpoints which contain other basic types in their typeset may only be used if these basic types are assumed to be instantiated in the musical surface being predicted or have already been predicted in an earlier stage.

The conversion is achieved by a function which maps elements of $[\tau]$ onto elements of $[\tau_b]$:

$$\Psi'_\tau : \xi^* \times [\tau] \rightarrow 2^{[\tau_b]}.$$

The function Ψ'_τ is implemented by creating a set of events each of which corresponds to a distinct basic element in $[\tau_b]$. A set of sequences is created by appending each of these events to the sequence of previously processed events in the composition. By calling the function Ψ_τ on each of these sequences each element in $[\tau_b]$ is put into the mapping with the current element of $[\tau]$. The mapping is, in general, many-to-one since the derived sequence $\Phi_\tau(e_1^j)$ could

represent many sequences of events other than e_1^j. As a result, the probability estimate returned by the model for the derived sequence must be divided equally amongst the basic event sequences onto which it maps.

A model m_τ must return a complete distribution over the basic attributes in $\langle\tau\rangle$. This does not present problems for basic viewpoints where the viewpoint domain is predefined to be the set of viewpoint elements occurring in the corpus. However, for derived viewpoints, such as `cpint`, it may not be possible to derive a complete distribution over [`cpitch`] from the set of derived elements occurring in the corpus. To address this problem, the domain of each derived type τ is set prior to prediction of each event such that there is a one-to-one correspondence between $[\tau]$ and the domain of the basic type $\tau_b \in \langle\tau\rangle$ currently being predicted. It is assumed that the modelling technique has some facility for assigning probabilities to events that have never occurred before (see §6.2.3). If no viewpoints predict some basic attribute then the *completion* of that attribute must be achieved on the basis of information from other sources or on the basis of a uniform distribution over the attribute domain. In the present research, m_{τ_b} was used to achieve the completion of attribute τ_b in such cases.

Once the distributions generated by each model in a multiple viewpoint system have been converted to complete distributions over the domain of a basic type, the distributions may be combined into final distributions for each basic type. The objective of the first experiment presented in this chapter was to empirically examine methods for achieving this combination.

7.3 Combining Viewpoint Prediction Probabilities

In this section, a novel technique is applied to the problem of combining the distributions generated by statistical models for different viewpoints. Let τ_b be the basic viewpoint currently under consideration and $[\tau_b]$ its domain. A multiple viewpoint system has n viewpoints $\tau_1, \ldots, \tau_n$, all which are derived from τ_b and whose type sets contain τ_b, and there exist corresponding sets of long-term models $LTM = \{ltm_1, ltm_2, \ldots, ltm_n\}$ and short-term models $STM = \{stm_1, stm_2, \ldots, stm_n\}$. A function is required which combines the distributions over $[\tau_b]$ generated by sets of models. As described in §7.2.1, this function is used in the first stage of prediction to combine the distributions generated by the LTM and the STM separately and, in the second stage of prediction, to combine the two combined distributions resulting from the first stage. Here, we employ an anonymous set of models $M = m_1, m_2, \ldots, m_n$ for the purposes of illustration.

In §6.2.4, a weighted arithmetic scheme was used to combine the predictions of the long- and short-term models in a single viewpoint system modelling `cpitch`. Conklin & Witten (1995) describe two other schemes for combining model predictions: the first converts the distributions into ranked lists, combines the rankings and transcribes the combined ranked list back into a probability distribution; the second method is based on the Dempster-Shafer theory of evidence and has been used for combining the predictions of long- and short-term models "with some success" (Conklin & Witten, 1995, p. 61). A novel method for combining the distributions generated by statistical models is presented here. The method uses a weighted geometric mean for combining individual probabilities which may then be applied to sorted distributions over $[\tau_b]$.[1]

A simple geometric mean of the estimated probabilities for each symbol $t \in [\tau_b]$ is calculated as:

$$p(t) = \frac{1}{R} \left(\prod_{m \in M} p_m(t) \right)^{\frac{1}{|M|}}$$

where R is a normalisation constant such that the resulting distribution over $[\tau_b]$ sums to unity. As in the case of the arithmetic mean, this technique may be improved by weighting the contributions made by each of the models such that:

$$p(t) = \frac{1}{R} \left(\prod_{m \in M} p_m(t)^{w_m} \right)^{\frac{1}{\sum_{m \in M} w_m}} \tag{7.1}$$

where R is a normalisation constant such that the resulting distribution over $[\tau_b]$ sums to unity. The entropy-based weighting technique described in the context of weighted arithmetic combination in §6.2.4 (Equation 6.9) may be used here to weight the contributions made by each model to the final estimates. As with weighted arithmetic combination, when the bias parameter to the weighting function $b = 0$, the weighted geometric scheme is equivalent to its unweighted counterpart since all models are assigned an equal weight of one (see Equation 6.9).

[1] In the present research, combination schemes based on the arithmetic mean are referred to as arithmetic combination and those based on the geometric mean as geometric combination. Similar distinctions have been made in the literature between linear and logarithmic opinion pools (Genest & Zidek, 1986), combining classifiers by averaging and multiplying (Tax *et al.*, 2000) and mixtures and products of experts (Hinton, 1999, 2000).

Hinton (1999, 2000) introduced the *products of experts* architecture for density estimation, where individual expert probabilities are multiplied and renormalised (using the *unweighted* geometric mean shown in Equation 7.1) as an alternative to combining probabilities using a *mixture* (a weighted arithmetic mean). Hinton argues that combining distributions through multiplication has the attractive property of making distributions "sharper" (or less uniform) than the component distributions. For a given element of the distributions it suffices for just one model to correctly assign that element a low estimated probability for the combined distribution to assign that element a low probability regardless of whether other models incorrectly assign that element a high estimated probability. Arithmetic combination, on the other hand, will tend to produce combined distributions that are more uniform than the component distributions and is prone to erroneously assigning relatively high estimated probabilities to irrelevant elements. However, since the combined distribution cannot be sharper than any of the component distributions arithmetic combination has the desirable effect of suppressing estimation errors (Tax *et al.*, 2000). Tax *et al.* (2000) examine the performance of unweighted arithmetic and geometric combination schemes in the context of multiple classifier systems. In accordance with theoretical predictions, an arithmetic scheme performs better when the classifiers operate on identical data representations and a geometric scheme performs better when the classifiers employ independent data representations.

On the basis of these theoretical and empirical considerations, it is predicted that the geometric combination scheme will outperform the arithmetic scheme for viewpoint combination where the models are derived from distinct data representations. Consider movement to a non scale degree as an example: a model associated with `cpitch` might return a high probability estimate for such a transition whereas a model associated with `cpintfref` is likely to return a low estimated probability. In cases such as this, it is preferable to trust the model operating over the more specialised data representation (*i.e.,* `cpintfref`). In the case of LTM-STM combination, however, it is predicted that the importance of suppressing estimation errors will outweigh the importance of trusting the estimates of one particular model. As a consequence, arithmetic combination is expected to outperform geometric combination in this second stage of combination. As an example, the LTM and STM will return low and high estimates, respectively, for n-grams which are common in the current composition but rare in the corpus as a whole. In cases such as this, it is preferable to suppress the estimation errors yielded by the LTM.

τ_b = cpitch		τ_1 = contour					τ_2 = cpintfref					
$[\tau_b]$	$H_{max}([\tau_b])$	$[\tau_1]$	$p_{\tau_1}([\tau_1])$	$p_{\tau_1}([\tau_b])$	H_{τ_1}	w_{τ_1}	$[\tau_2]$	$p_{\tau_2}([\tau_2])$	$p_{\tau_2}([\tau_b])$	H_{τ_2}	w_{τ_2}	$p_{combined}([\tau_b])$
78	3.585			0.049	3.373	1.063	11	0.011	0.011	2.392	1.499	0.014
77	3.585			0.049	3.373	1.063	10	0.038	0.038	2.392	1.499	0.042
76	3.585			0.049	3.373	1.063	9	0.005	0.005	2.392	1.499	0.016
75	3.585	1	0.294	0.049	3.373	1.063	8	0.008	0.008	2.392	1.499	0.021
74	3.585			0.049	3.373	1.063	7	0.195	0.195	2.392	1.499	0.135
73	3.585			0.049	3.373	1.063	6	0.005	0.005	2.392	1.499	0.016
72	3.585	0	0.235	0.235	3.373	1.063	5	0.121	0.121	2.392	1.499	0.196
71	3.585			0.094	3.373	1.063	4	0.394	0.394	2.392	1.499	0.267
70	3.585			0.094	3.373	1.063	3	0.196	0.196	2.392	1.499	0.177
69	3.585	-1	0.471	0.094	3.373	1.063	2	0.021	0.021	2.392	1.499	0.048
68	3.585			0.094	3.373	1.063	1	0.005	0.005	2.392	1.499	0.021
67	3.585			0.094	3.373	1.063	0	0.021	0.021	2.392	1.499	0.048

Table 7.1: An illustration of the weighted geometric scheme for combining the predictions of different models; a bias value of $b = 1$ is used in calculating model weights and all intermediate calculations are made on floating point values rounded to 3 decimal places.

In order to illustrate the mechanics of the weighted geometric combination scheme, consider the example of predicting the pitch of the final event in the melodic phrase shown in Figure 7.2 using two long-term models associated with the types `contour` and `cpintfref` respectively. Each model has already been trained (using the methods described in Chapter 6) on preprocessed sequences of the appropriate type (see §7.2.2) and the basic type of interest τ_b = `cpitch`. As shown in the first column of Table 7.1, our assumed pitch alphabet will be all 12 chromatic pitches between G_4 and $F\sharp_5$: $[\texttt{cpitch}] = \{67, \ldots, 78\}$. Following Equation 6.4, the maximum entropy of a probability distribution constructed over this alphabet $H_{max}([\tau_b]) = 3.585$, as shown in the second column of Table 7.1; this quantity will be needed to calculate the weights assigned to each model.

The next five columns in Table 7.1 concern the model for the attribute τ_1 = `contour` whose domain $[\texttt{contour}] = \{-1, 0, 1\}$ (see §5.4.1). In the example shown in Table 7.1, the model returns estimated probabilities of 0.294, 0.471 and 0.235 ($p_{\tau_1}([\tau_1])$) for rising, falling and stationary pitch contours to the final event of the phrase (using the methods discussed in Chapter 6). Since rising and falling contours could each correspond to a number of distinct pitches, this distribution is converted to a distribution over the basic pitch alphabet ($p_{\tau_1}([\tau_b])$) by dividing the estimates for rising and falling contour equally amongst the set of corresponding elements of [`cpitch`]. The entropy of this distribution, H_{τ_1}, is calculated according to Equation 6.3. Finally, the weight associated with this model is calculated from H_{τ_1} and $H_{max}([\tau_b])$ using Equation 6.9; a bias value of $b = 1$ has been used in this example.

As shown in the next five columns in Table 7.1, an analogous process is carried out for the model associated with the type τ_2 = `cpintfref`, the major difference being that the completion of the estimates for the basic type (`cpitch`) is unnecessary since, in this example, there is a one-to-one correspondence between [`cpitch`] and [`cpintfref`]. As a consequence, the columns in Table 7.1 representing the estimates over the derived alphabet ($p_{\tau_2}([\tau_2])$) and the basic alphabet ($p_{\tau_2}([\tau_b])$) are equivalent. Table 7.1 now contains all the information needed to derive the final combined estimates for each element of [`cpitch`]. This is achieved by using Equation 7.1 to weight each model's estimates, multiply the weighted estimates in each row, and then normalise such that the resulting distribution over [`cpitch`] sums to unity. The resulting combined estimates are shown in the final column of Table 7.1.

7.4 Experimental Methodology

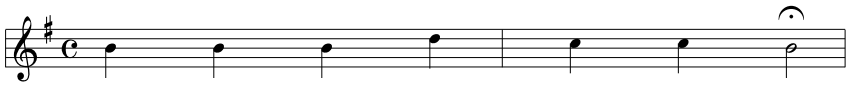
The corpus of music used in this experiment was Dataset 2 consisting of 185 of the chorale melodies harmonised by J. S. Bach (see Chapter 4). The evaluation metric for model performance was cross entropy, as defined in Equation 6.6, computed by 10-fold cross-validation (Dietterich, 1998; Mitchell, 1997) over this corpus. The statistical model used was a smoothed n-gram model described as LTM+C*I—STMC*UI within the model syntax defined in §6.3. While the LTM+C*I—STMC*I model was found to be the best-performing predictor of `cpitch` in Chapter 6, the LTM+C*I—STMX*UI model performed almost as well. In preliminary pilot experiments with multiple viewpoint models, the latter model consistently outperformed the former (albeit by a small margin). Unless otherwise specified a LTM+C*I—STMX*UI has been used in the remainder of the present research.

The goal of the first experiment was to examine methods for combining viewpoint predictions and, accordingly, a constant set of viewpoints was used corresponding to the best performing of the multiple viewpoint systems described by Conklin & Witten (1995). This system consists of the following viewpoints:

$$\texttt{cpintfref} \otimes \texttt{cpint},$$
$$\texttt{cpint} \otimes \texttt{ioi},$$
$$\texttt{cpitch},$$
$$\texttt{cpintfref} \otimes \texttt{fib}$$

and is capable of modelling the basic type `cpitch` alone. The first of the component viewpoints of this system represents a link between scale degree and pitch interval, the second a link between pitch interval and inter-onset interval, the third chromatic pitch and the fourth a link between scale degree and a test type indicating whether or not an event is the first in the current bar. See Tables 5.2 and 5.4 for details of each of the viewpoints in this system and Figure 7.2 for an example use of these viewpoints in representing an excerpt from a chorale melody in terms of viewpoint sequences.

The experiment compared the performance of the weighted arithmetic combination scheme described in §6.2.4 with that of the weighted geometric combination scheme described in §7.3 in both stages of combination with the bias settings drawn from the set {0,1,2,3,4,5,6,7,8,16,32}. In preliminary experiments, the Dempster-Shafer and rank-based combination schemes examined



`onset`	0	24	48	72	96	120	144
`cpitch`	71	71	71	74	72	72	71
`ioi`	⊥	24	24	24	24	24	24
`fib`	T	F	F	F	T	F	F
`cpint`	⊥	0	0	3	-2	0	-1
`cpintfref`	4	4	4	7	5	5	4
`cpint`⊗`ioi`	⊥	(0 24)	(0 24)	(3 24)	(-2 24)	(0 24)	(-1 24)

Figure 7.2: The first phrase of the melody from Chorale 151 *Meinen Jesum laß' ich nicht, Jesus* (BWV 379) represented as viewpoint sequences in terms of the component viewpoints of the best-performing system reported by Conklin & Witten (1995).

by Conklin & Witten (1995) were found to perform less well than these two methods (when optimally weighted) and therefore were not included in the experiment reported here. On the basis of the empirical results and theoretical considerations reviewed in §7.3, it is predicted that the geometric combination scheme will yield performance improvements over the arithmetic combination scheme particularly when combining models associated with different viewpoints.

An obvious limitation of the first experiment is that it is restricted to a single set of viewpoints. Accordingly, a second experiment was run to examine the performance of different multiple viewpoint systems and, in particular, to discover those which are capable of reducing model entropy still further. In this experiment, the combination methods used were those that yielded best performance in the first experiment (*i.e.,* geometric combination with a bias of 7 for LTM-STM combination and a bias of 2 for Viewpoint combination).

The selection of viewpoints to reduce model entropy can be viewed as a problem of *feature selection* where the goal is to attempt to reduce the number of dimensions considered in a task so as to improve performance according to some evaluation function (Aha & Bankert, 1996). Computational methods for feature selection typically consist of an algorithm which searches the space of feature subsets and an evaluation function which returns a performance measure associated with each feature subset. The goal is to search the space of feature subsets in order to maximise this measure. The performance measure used in this experiment was cross entropy, as defined in Equation 6.6, computed by 10-fold cross-validation over Dataset 2 (see Chapter 4). The feature sets used in these experiments consist of subsets of the attribute types shown in Table 5.2 and Table 5.4. Note that the product types used in the search cor-

respond to a selected subset of the space of possible product types (see §5.4.4). In previous research, Hall (1995) has used a genetic algorithm for selecting sets of viewpoints in order to minimise model entropy over a held-out test set.

The simple sequential search algorithm, known as *stepwise selection*, employed in the current research is commonly used both in machine learning (Aha & Bankert, 1996; Blum & Langley, 1997; Kohavi & John, 1996) and statistical analysis (Krzanowski, 1988). Given an initial set of features, the algorithm considers on each iteration all single feature additions and deletions from the current feature set and selects the addition or deletion that yields the most improvement in the performance measure. Following the methodological principle of Ockham's razor, the algorithm selects the optimal deletion if one exists before considering any additions. The algorithm terminates when no single addition or deletion yields an improvement. *Forward stepwise selection* corresponds to the situation where the initial set of features is the empty set, whilst *backward stepwise elimination* corresponds to the situation where the algorithm is initialised with the full set of features (John *et al.*, 1994). A forward stepwise selection algorithm has been used for feature selection in this experiment. Given n features, the size of the space of feature subsets is 2^n. The forward stepwise selection algorithm, on the other hand, is guaranteed to terminate in a maximum of n^2 iterations (John *et al.*, 1994). However, while the solution returned will be locally optimal, it is not guaranteed to be globally optimal.

7.5 Results and Discussion

7.5.1 Model Combination

The results of the first experiment are shown in Table 7.2 in which the columns represent the settings for viewpoint combination and the rows indicate the settings for LTM-STM combination. The results are given in terms of cross entropies for each combination of settings for viewpoint and LTM-STM combination. Table 7.2 is divided into four sections corresponding to the use of arithmetic or geometric methods for viewpoint or LTM-STM combination. Figures in bold type represent the lowest entropies in each of the four sections of the table. The results are also plotted graphically in Figure 7.3. The first point to note is that the multiple viewpoint system is capable of predicting the dataset with much lower entropies (*e.g.,* 2.045 bits/symbol) than those reported in Chapter 6 for a system modelling chromatic pitch alone (*e.g.,* 2.342 bits/symbol) on the same corpus. A paired t test confirms the significance of this difference $[t(184) = 15.714, p < 0.001]$.

This test was carried out for the paired differences between the entropy estimates obtained using the two models for each of the 185 chorale melodies in the dataset. However, since the estimates within each 10-fold partition share the same training set, they are not independent and, therefore, violate one of the assumptions of the t test. In order to address this issue, a second paired t test was carried out to compare the paired estimates averaged for each of the 10 resampling partitions. This follows the procedure adopted in §6.4.5 although since only one dataset is used, the test has a small number of degrees of freedom and is offered as a supplement to (rather than a replacement for) the previous test. Nonetheless, the test confirmed the significance of the difference between the entropy estimates of the two models [$t(9) = 24.09, p < 0.001$].[2]

Overall, these results replicate the findings of Conklin & Witten (1995) and lend support to their assertion that the multiple viewpoint framework can increase the predictive power of statistical models of music. It is also clear that the use of an entropy-based weighting scheme improves performance and that performance can be further improved by tuning the bias parameter which gives exponential bias towards models with lower relative entropies (Conklin, 1990).

Regarding the combination methods, the results demonstrate, as predicted, that the weighted geometric combination introduced in §7.3 tends to outperform arithmetic combination and that this effect is much more marked in the case of viewpoint combination than it is for LTM-STM combination. This finding corroborates theoretical predictions and empirical results derived from the literature on combining classifier systems (see §7.3). When combining viewpoint predictions (derived from distinct data representations), a geometric scheme performs better since it trusts specialised viewpoints to correctly assign low probability estimates to certain elements (Hinton, 1999, 2000; Tax *et al.*, 2000).

[2] Since the training sets show significant overlap between the k-fold partitions, these estimates are not strictly independent either (Dietterich, 1998). Difficulties such as this are an inevitable consequence of using a machine learning approach with small datasets (Dietterich, 1998; Mitchell, 1997).

		Viewpoint Combination																					
		Arithmetic											Geometric										
		0	1	2	3	4	5	6	7	8	16	32	0	1	2	3	4	5	6	7	8	16	32
A	0	2.493	2.437	2.393	2.363	2.342	2.327	2.316	2.309	2.304	2.290	2.291	2.357	2.321	2.299	2.286	2.278	2.274	2.271	2.270	2.269	2.274	2.285
r	1	2.434	2.368	2.317	2.281	2.257	2.241	2.230	2.222	2.217	2.207	2.212	2.256	2.216	2.192	2.180	2.175	2.173	2.173	2.174	2.175	2.188	2.203
i	2	2.386	2.317	2.264	2.229	2.207	2.193	2.184	2.178	2.175	2.171	2.178	2.189	2.150	2.130	2.123	2.122	2.124	2.126	2.130	2.133	2.152	2.169
t	3	2.350	2.279	2.228	2.196	2.177	2.166	2.160	2.156	2.155	2.159	2.168	2.146	2.111	2.096	2.094	2.097	2.102	2.107	2.112	2.117	2.142	2.160
h	4	2.323	2.253	2.204	2.175	2.159	2.150	2.147	2.145	2.146	2.157	2.169	2.119	2.088	2.077	2.078	2.085	2.092	2.100	2.107	2.113	2.142	2.161
m	5	2.303	2.234	2.188	2.161	2.147	2.141	2.139	2.140	2.141	2.158	2.173	2.102	2.074	2.066	2.070	2.079	2.089	2.098	2.106	2.113	2.146	2.167
e	6	2.288	2.221	2.176	2.152	2.140	2.136	2.135	2.137	2.140	2.161	2.179	2.091	2.066	2.060	2.066	2.077	2.088	2.098	2.108	2.116	2.151	2.174
t	7	2.276	2.211	2.168	2.146	2.136	2.133	2.134	2.136	2.140	2.165	2.184	2.085	2.061	2.057	2.064	2.076	2.088	2.099	2.110	2.118	2.156	2.180
i	8	2.268	2.204	2.163	2.142	2.133	**2.131**	2.133	2.136	2.140	2.168	2.189	2.080	2.057	2.055	2.064	2.076	2.089	2.101	2.112	2.121	2.161	2.186
c	16	2.243	2.186	2.152	2.136	2.132	2.133	2.138	2.143	2.149	2.184	2.212	2.073	**2.053**	2.054	2.066	2.081	2.097	2.111	2.123	2.134	2.182	2.212
	32	2.239	2.185	2.154	2.140	2.138	2.140	2.145	2.151	2.157	2.195	2.226	2.074	2.055	2.058	2.070	2.086	2.103	2.118	2.132	2.143	2.194	2.226
	0	2.496	2.437	2.386	2.346	2.316	2.294	2.278	2.266	2.257	2.237	2.240	2.311	2.267	2.238	2.222	2.213	2.208	2.207	2.206	2.207	2.219	2.234
G	1	2.425	2.354	2.295	2.252	2.222	2.202	2.188	2.178	2.172	2.160	2.165	2.200	2.155	2.129	2.118	2.114	2.114	2.116	2.119	2.122	2.141	2.157
e	2	2.372	2.298	2.240	2.201	2.176	2.161	2.151	2.145	2.142	2.142	2.150	2.138	2.098	2.081	2.077	2.079	2.084	2.090	2.096	2.101	2.126	2.143
o	3	2.334	2.260	2.206	2.172	2.152	2.141	2.135	2.133	2.132	2.141	2.154	2.104	2.070	2.059	2.060	2.067	2.076	2.084	2.092	2.099	2.13	2.149
m	4	2.307	2.235	2.185	2.155	2.139	2.131	2.128	2.128	2.129	2.146	2.163	2.086	2.057	2.050	2.054	2.064	2.075	2.085	2.095	2.103	2.138	2.159
e	5	2.288	2.218	2.171	2.145	2.132	2.127	2.126	2.127	2.130	2.152	2.171	2.077	2.051	2.046	2.053	2.065	2.077	2.089	2.099	2.108	2.146	2.169
t	6	2.275	2.207	2.163	2.139	2.129	**2.125**	2.126	2.128	2.132	2.158	2.179	2.072	2.048	**2.045**	2.054	2.067	2.080	2.092	2.103	2.113	2.154	2.178
r	7	2.265	2.200	2.158	2.136	2.127	**2.125**	2.127	2.130	2.134	2.163	2.186	2.069	2.047	**2.045**	2.055	2.069	2.083	2.096	2.107	2.117	2.160	2.185
i	8	2.258	2.194	2.155	2.134	2.127	**2.125**	2.128	2.132	2.136	2.167	2.192	2.068	2.047	2.046	2.057	2.071	2.085	2.099	2.111	2.121	2.165	2.191
c	16	2.240	2.184	2.151	2.136	2.132	2.133	2.138	2.144	2.150	2.186	2.216	2.070	2.051	2.053	2.065	2.081	2.098	2.112	2.125	2.136	2.186	2.217
	32	2.239	2.185	2.154	2.141	2.138	2.141	2.146	2.151	2.158	2.198	2.229	2.073	2.055	2.057	2.070	2.087	2.104	2.12	2.134	2.145	2.197	2.230

Table 7.2: The performance on Dataset 2 of models using weighted arithmetic and geometric combination methods with a range of bias settings.

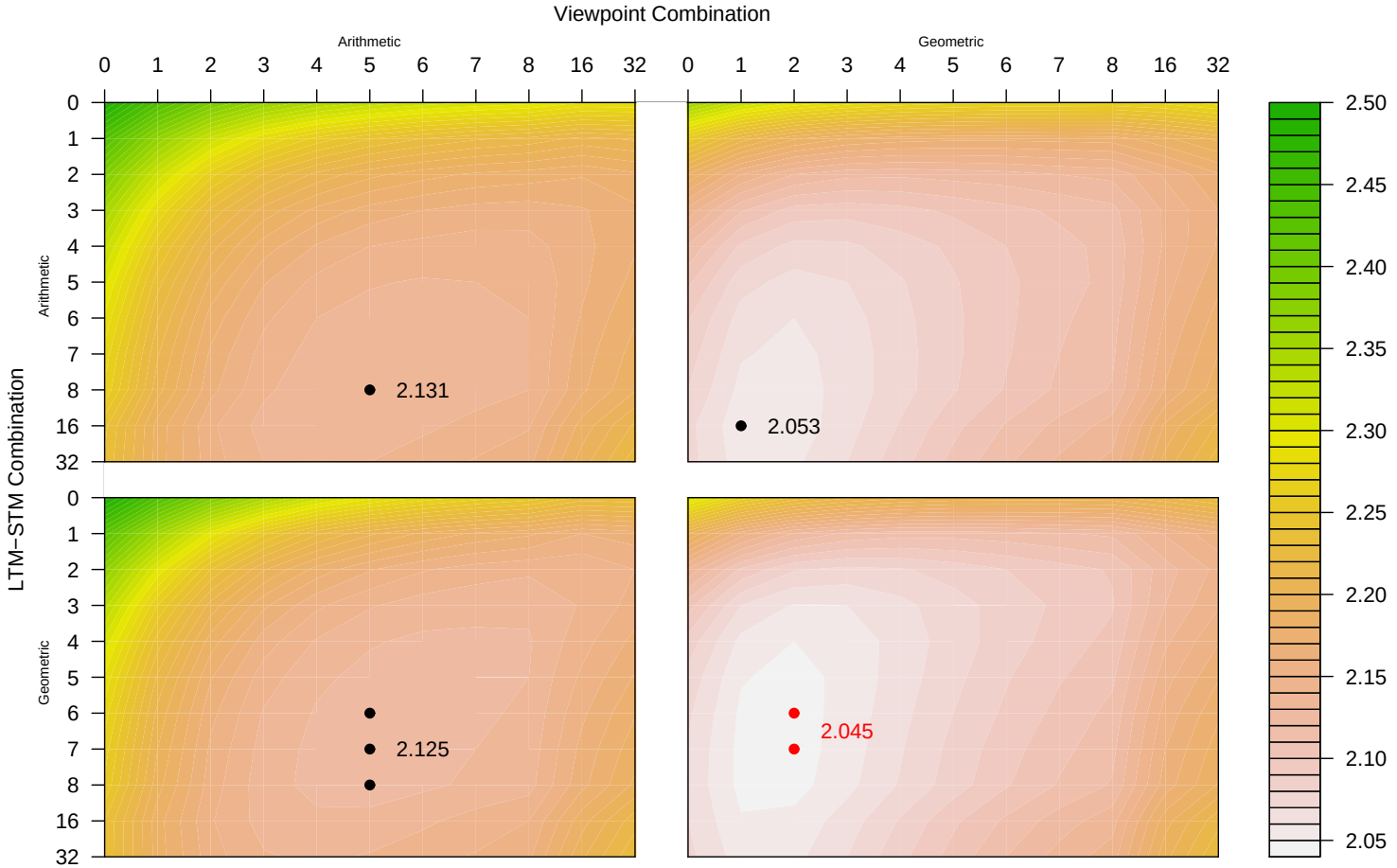


Figure 7.3: The performance on Dataset 2 of models using weighted arithmetic and geometric combination methods with a range of bias settings.

When combining LTM-STM predictions (where each distribution is already the result of combining the viewpoint predictions), on the other hand, a premium is placed on minimising estimation errors (Hinton, 1999, 2000; Tax *et al.*, 2000). The finding that geometric combination still outperforms arithmetic combination may indicate that this suppression of errors is not necessary perhaps because the LTM itself induces local structure in the current melody by adding n-grams online as prediction progresses just as the STM does (see Chapter 6). Finally, it is possible that the difference in relative performance of the geometric and arithmetic schemes for LTM-STM and viewpoint combination is a result of the order in which these combinations are performed (see Figure 7.1). However, it is hypothesised that this is not the case and the observed pattern of results arises from the difference between combining distributions derived from distinct data representations as opposed to combining two distributions already combined from the same sets of representations. Further research is required to examine these hypotheses in more depth.

Another aspect of the results that warrants discussion is the effect on performance of the bias parameter which gives an exponential bias towards distributions with lower relative entropy. Overall performance seems to be optimised when the bias for LTM-STM combination is relatively high (between 6 and 16) and the bias for viewpoint combination is relatively low (between 1 and 5). It seems likely that this is due, in part, to the fact that at the beginning of a composition, the STM will generate relatively high entropy distributions due to lack of context. In this case, it will be advantageous for the system to strongly bias the combination towards the LTM predictions. This is not an issue when combining viewpoint predictions and more moderate bias values tend to be optimal. Other research has also found that high bias values for the combination of the LTM-STM predictions tend to improve performance leading to the suggestion that the weight assigned to the STM could be progressively increased from an initially low value at the beginning of a composition as more events are processed (Conklin, 1990).

The results shown in Table 7.2 also reveal an inverse relationship between the optimal bias settings for LTM-STM combination and those for viewpoint combination. With high bias values for LTM-STM combination, low bias values for viewpoint combination tend to be optimal and *vice versa*. High bias settings will make the system bolder in its estimation by strongly favouring sharper distributions while low bias settings will lead it to more conservative predictions. On these grounds, with all other things being equal, we would expect moderate bias values to yield optimal performance. If an extreme bias setting is preferred

Stage	Viewpoint Added	H
1	cpint⊗dur	2.214
2	cpintfref⊗cpintfip	2.033
3	cpitch⊗dur	1.991
4	cpintfref⊗fib	1.973
5	thrtactus	1.966
6	cpintfref⊗dur	1.963
7	cpint⊗dur-ratio	1.960
8	cpintfip	1.955
9	thrfiph	1.953

Table 7.3: The results of viewpoint selection for reduced entropy over Dataset 2.

in one stage of combination for some other reason (*e.g.,* the case of LTM-STM combination just discussed), the negative effects may, it seems, be counteracted to some extent by using settings at the opposing extreme in the other stage. Although these arguments are general, we would expect the optimal bias settings themselves to vary with different data, viewpoints and predictive systems.

7.5.2 Viewpoint Selection

The results of viewpoint selection are shown in Table 7.3 which gives the entropies obtained with the optimal multiple viewpoint systems selected at each stage. In all stages of selection, viewpoints were added to the optimal system; removing viewpoints failed to reduce model uncertainty at any stage. The first point to note about the results is that viewpoint selection results in a multiple viewpoint system exhibiting lower model uncertainty than the system used in the first experiment (see §7.5.1) in which the component viewpoints were hand-selected by Conklin & Witten (1995). Thus, the average cross entropy of the data given the optimal multiple viewpoint system derived in this experiment was 1.953 bits/symbol compared with 2.045 bits/symbol obtained in the first experiment. The significance of this difference was confirmed by paired t tests over all 185 chorale melodies [$t(184) = 7.810, p < 0.001$] and averaged for each 10-fold partition of the dataset [$t(9) = 10.701, p < 0.001$] (see §7.5.1).

A number of interesting observations can be made regarding the actual viewpoints selected in this experiment. First, the multiple viewpoint system selected is dominated by linked and threaded viewpoints; only one primitive type, cpintfip, was selected and only relatively late in the selection process.[3]

[3]Threaded models are often referred to as *long distance n-grams* in research on statistical language modelling where they have been shown to improve model performance (Huang *et al.*,

Many of the linked viewpoints selected represent conjunctions of pitch and time related attribute types. This suggests that strong correlations exist in the corpus of chorale melodies between pitch structure and rhythmic structure. It is perhaps not surprising that the majority of the viewpoints selected model relative pitch structure (*e.g.,* `cpint`, `cpintfref` and `cpintfip`) rather than pitch height. This suggests that regularities in melodic structure within the corpus tend to be expressed in terms of pitch intervals or defined in relation to significant tones. In addition, the selection of `thrtactus` suggests that stylistic commonalities in interval structure can be found when tones occurring on weak beats (*e.g.,* passing notes) are ignored. Finally, the selection of `thrfiph` (and the failure to select `thrbar`) is likely to reflect the relative importance of phrase structure over higher level metric structure in this corpus.

7.6 Summary

In this chapter, the predictive system developed in Chapter 6 has been applied within the multiple viewpoint representational framework presented in Chapter 5. A novel combination technique based on a weighted geometric mean was introduced in §7.3 and empirically compared with an existing technique based on a weighted arithmetic mean. The entropy-based technique described in §6.2.4 was used to compute the combination weights. This method accepts a parameter which fine-tunes the exponential bias given to distributions with lower relative entropy. In the experiment, a range of parameterisations of the two techniques were evaluated using cross entropy computed by 10-fold cross-validation over Dataset 2 (see Chapter 4).

The results presented in §7.5 demonstrate that the weighted geometric combination introduced in the present research tends to outperform arithmetic combination especially for the combination of viewpoint models. Drawing on related findings in machine learning research on combining multiple classifiers, it was hypothesised that this asymmetry arises from the difference between combining distributions derived from distinct data representations as opposed to combining distributions derived from the same data representations. In a second experiment, a feature selection algorithm was applied to select multiple viewpoint systems with lower cross entropy over Dataset 2. The uncertainty associated with the resulting system was significantly lower than that of the multiple viewpoint system used in the first experiment. The selected viewpoints highlight some interesting stylistic regularities of the corpus.

1993; Mahajan *et al.*, 1999; Simons *et al.*, 1997).

CHAPTER 8

MODELLING MELODIC EXPECTANCY

8.1 Overview

The objective in this chapter is to examine the predictive system developed in Chapters 6 and 7 as a model of *melodic expectancy* in human perception of music. The concept of expectancy has long been of interest in psychology and the cognitive sciences. Expectancy is simply the anticipation of forthcoming events based on currently available information and may vary independently in both strength and specificity. Following Eerola (2004b), the term *expectancy* is used here to refer to a generic state of anticipation of forthcoming events and the term *expectation* to refer, more specifically, to the anticipation of a particular future event or events. The ability to anticipate forthcoming events is clearly very important in an adaptive sense since it may aid the direction of attention to, and rapid processing of, salient environmental stimuli as well as facilitate the preparation and execution of appropriate responses to them. In addition, conflicts or correspondences between actual and anticipated effects often entail significant psychological and biological effects.

The generation of expectations is recognised as being an especially important factor in music cognition. From a music-analytic perspective, it has been argued that the generation and subsequent confirmation or violation of expectations is critical to aesthetic experience and the communication of emotion and meaning in music (Meyer, 1956; Narmour, 1990). From a psychological perspective, expectancy has been found to influence recognition memory for music (Schmuckler, 1997), the production of music (Carlsen, 1981; Schmuck-

ler, 1989, 1990; Thompson *et al.*, 1997; Unyk & Carlsen, 1987), the perception of music (Cuddy & Lunny, 1995; Krumhansl, 1995a,b; Krumhansl *et al.*, 1999, 2000; Schellenberg, 1996, 1997; Schellenberg *et al.*, 2002; Schmuckler, 1989) and the transcription of music (Unyk & Carlsen, 1987). While most empirical research has examined the influence of melodic structure, expectancy in music also reflects the influence of rhythmic and metric structure (Jones, 1987; Jones & Boltz, 1989) and harmonic structure (Bharucha, 1987; Schmuckler, 1989). Patterns of expectation may be influenced both by intra-opus memory for specific musical structures as well as by more abstract extra-opus schemata acquired through extensive exposure to music (Bharucha, 1987; Krumhansl *et al.*, 1999, 2000; Narmour, 1990).

The research presented in this chapter examines the cognitive mechanisms underlying melodic expectations. Narmour (1990, 1992) has proposed a detailed and influential theory of expectancy in melody which attempts to characterise the set of implied continuations to an incomplete melodic pattern. According to the theory, the expectations of a listener are influenced by two distinct cognitive systems: first, a bottom-up system consisting of Gestalt-like principles which are held to be innate and universal; and second, a top-down system consisting of style-specific influences on expectancy which are acquired through extensive exposure to music in a given style. Krumhansl (1995b) has formulated the bottom-up system of the IR theory as a quantitative model, consisting of a small set of symbolic rules, which is summarised in §8.2.2 in terms of its principal characteristics and the manner in which it differs from the IR theory of Narmour (1990). This model has formed the basis of a series of empirical studies, reviewed in §8.2.3, which have examined the degree to which the expectations of listeners conform to the predictions of the IR theory and have led to several different formulations of the principles comprising the bottom-up component of the model.

While this body of research suggests that the expectations of listeners in a given experiment may be accounted for by some collection of principles intended to reflect the bottom-up and top-down components of Narmour's theory, the present research is motivated by empirical data that question the existence of a small set of universal bottom-up rules that determine, in part, the expectations of a listener. According to the theory presented in §8.3.1, expectancy in melody can be accounted for entirely in terms of the induction of statistical regularities in sequential melodic structure without recourse to an independent system of innate symbolic predispositions. While innate constraints on music perception certainly exist (Justus & Hutsler, 2005; McDermott & Hauser, 2005;

Trehub, 1999), it is argued here that they are unlikely to be found in the form of rules governing sequential dependencies between musical events. According to the account developed here, observed patterns of melodic expectation can be accounted for in terms of the induction of statistical regularities existing in the music to which the listener is exposed. Patterns of expectation that do not vary between musical styles are accounted for in terms of simple regularities in music whose ubiquity may be related to the constraints of physical performance. If this is the case, there is no need to make additional (and problematic) assumptions about innate representations of sequential dependencies between perceived events (Elman *et al.*, 1996).

The specific goals of this research are twofold. The first is to examine whether models of melodic expectancy based on statistical learning are capable of accounting for the patterns of expectation observed in empirical behavioural research. If such models can account for the behavioural data as well as existing implementations of the IR theory, there would be no need to invoke symbolic rules as universal properties of the human cognitive system. To the extent that such models can be found to provide a more powerful account of the behavioural data, the IR theory (as currently implemented) may be viewed as an inadequate cognitive model of melodic expectancy by comparison. Instead of representing innate and universal constraints of the perceptual system, the bottom-up principles may be taken to represent a formalised approximate description of the mature behaviour of a cognitive system of inductive learning. The second goal of the present research is to undertake a preliminary examination of the kinds of melodic feature that afford regularities capable of supporting the acquisition of the patterns of expectation exhibited by listeners.

In order to achieve these goals, the statistical system developed in Chapters 6 and 7 is used to model empirically observed patterns of human expectation and the fit is compared to that obtained with a quantitative formulation of the IR theory consisting of two bottom-up principles (Schellenberg, 1997). The experimental methodology used to examine the behaviour of the statistical model is discussed in §8.4.

The question of distinguishing acquired and inherited components of behaviour is a thorny one, all the more so in relation to the perception of cultural artefacts (which are both created and appreciated through the application of the human cognitive system). Following Cutting *et al.* (1992), three criteria are used to compare the two cognitive models of melodic expectation. The first criterion is *scope*, which refers to the degree to which a theory accounts for a broad range of experimental data elicited in a variety of contexts. In order

to compare the scope of the two models, the extent to which they account for the patterns of expectation exhibited by listeners is examined and compared in three experiments presented in §8.5, §8.6 and §8.7. The three experiments examine expectations elicited in the context of increasingly complex melodic stimuli and also incorporate analyses of more detailed hypotheses concerning the melodic features that afford regularities capable of supporting the acquisition of the observed patterns of expectation.

The second criterion introduced by Cutting *et al.* (1992) is *selectivity*, which refers to the degree to which a theory accounts specifically for the data of interest and does not predict unrelated phenomena. In order to compare the models on the basis of selectivity, the ability of each model to account for random patterns of expectation is assessed and compared in each experiment.

The third criterion discussed by Cutting *et al.* (1992) is the *principle of parsimony* (or *simplicity*): a general methodological heuristic expressing a preference for the more parsimonious of two theories that each account equally well for observed data. Although the precise operational definition of parsimony is a point of debate in the philosophy of science, variants of the heuristic are commonly used in actual scientific practice (Nolan, 1997; Popper, 1959; Sober, 1981). This provides some evidence that the principle is normative; *i.e.,* that it actually results in successful theories. Further evidence along these lines is provided by the fact that simplicity is commonly used a heuristic bias in machine learning (Mitchell, 1997) and for hypothesis selection in abductive reasoning (Paul, 1993).

Furthermore, quantifying the principle of parsimony in terms of algorithmic information theory demonstrates that simple encodings of a set of data also provide the most probable explanations for that data (Chater, 1996, 1999; Chater & Vitányi, 2003). In the closely related field of Bayesian inference, it is common to compare models according to their simplicity, measured as a function of the number of free parameters they possess and the extent to which these parameters need to be finely tuned to fit the data (Jaynes, 2003; MacKay, 2003). Chater (1999) presents simplicity as a rational analysis of perceptual organisation on the basis of these normative justifications together with evidence that simple representations of experience are preferred in perception and cognition. Although this application of simplicity is not a primary concern in the present research, we touch on it again briefly in §8.4 as a justification for preferring small feature sets and when discussing the results of Experiment 3 in §8.7.2.

In psychology (as in many other scientific fields), the relative parsimony of comparable models is most commonly defined in terms of the number of free

parameters in each model (Cutting *et al.*, 1992). Here, however, we use the principle in a more general sense where the existence of a theoretical component assumed by one theory is denied leading to a simpler theory (Sober, 1981). To the extent that the theory of inductive learning is comparable to the top-down component of the IR theory (and in the absence of specific biological evidence for the innateness of the bottom-up principles), the former theory constitutes a more parsimonious description of the cognitive system than the latter since additional bottom-up constraints *assumed* to constitute part of the cognitive system are replaced by equivalent constraints *known* to exist in the environment. In order to test this theoretical position, we examine the extent to which the statistical model subsumes the function of the two-factor model of expectancy in accounting for the behavioural data in each experiment.

Finally, the chapter concludes in §8.8 with a general discussion of the experimental results, their implications and some promising directions for further development of the theory.

8.2 Background

8.2.1 Leonard Meyer's Theory of Musical Expectancy

In his book, *Emotion and Meaning in Music*, Meyer (1956) discusses the dynamic cognitive processes in operation when we listen to music and how these processes not only underlie the listener's understanding of musical structure but also give rise to the communication of affect and the perception of meaning in music. Broadly speaking, Meyer proposes that meaning arises through the manner in which musical structures activate, inhibit and resolve expectations concerning other musical structures in the mind of the listener. Meyer notes that expectations may differ independently in terms of the degree to which they are passive or active, their strength and their specificity. He contends, in particular, that affect is aroused when a passive expectation induced by antecedent musical structures is temporarily inhibited or permanently blocked by consequent musical structures. The perceptual uncertainty caused by such a violation of passive expectations may arise from different sources; it may depend on the listener's familiarity with a musical genre or a particular piece of music or the composer may deliberately introduce structures to violate the expectations of the listener for aesthetic effect (Meyer, 1957).

Meyer discusses three ways in which the listener's expectations may be violated. The first occurs when the expected consequent event is delayed, the second when the antecedent context generates ambiguous expectations about

consequent events and the third when the consequent is unexpected. While the particular effect of the music is clearly dependent on the strength of the expectation, Meyer argues that it is also conditioned by the specificity of the expectation. Meaning may be attributed to the antecedent structure as a consequence both of the expectations it generates and its relationship with the consequent structure once this is apprehended.

In his later work, Meyer (1973) conducted more detailed analyses of the *melodic structures* or *processes* in Western tonal music that give rise to more or less specific expectations in the listener. A *linear pattern*, for example, consists of a diatonic scale, a chromatic scale or some mixture of the two and creates an expectation for the pattern to continue in stepwise motion of seconds and thirds. A *gap-fill pattern*, on the other hand, consists of a large melodic interval (the *gap*) which creates an expectation for a succession of notes that fill the gap by presenting all or most of the notes skipped over by the gap.[1] Rosner & Meyer (1982, 1986) have provided some experimental support for the psychological validity of such melodic processes. Rosner & Meyer (1982) trained listeners to distinguish a number of passages of Western tonal music exemplifying either a gap-fill or a changing-note pattern. The subjects were subsequently able to classify correctly new instances of the two processes. Rosner & Meyer (1986) extended these findings by demonstrating that listeners rated passages of classical and romantic music based on the same melodic process as more similar to each other than passages based on different melodic processes. While von Hippel (2000a) has conducted a re-analysis of the data obtained by Rosner & Meyer (1982, 1986) which suggests that gap-fill patterns play little or no role in the classification tasks (see also §8.2.3.2), Schmuckler (1989) reports specific experimental evidence that listener's expectations follow the predictions of linear and gap-fill melodic processes.

8.2.2 The Implication-Realisation Theory

Narmour (1990, 1992) has extended Meyer's approach into a complex theory of melodic perception called the *Implication-Realisation* (IR) theory. The theory posits two distinct perceptual systems – the *bottom-up* and *top-down systems* of melodic implication. While the principles of the former are held to be hard-

[1] Other melodic processes discussed by Meyer (1973) are more complex. A *changing-note pattern*, for example, is one in which the main structural tones of the pattern consist of the tonic followed by the seventh and second scale degrees (in either order) followed by a return to the tonic. A *complementary pattern* is one in which a model pattern consisting of the main structural tones of a phrase is followed by a complementary model in which the direction of motion is inverted. Other melodic processes involve *Adeste Fideles patterns*, *triadic patterns* and *axial patterns* (see Rosner & Meyer, 1982, 1986, for a summary).

wired, innate and universal, the principles of the latter are held to be learnt and hence dependent on musical experience.

> The top-down system is flexible, variable and empirically driven ... In contrast, the bottom-up mode constitutes an automatic, unconscious, preprogrammed, "brute" system.
>
> (Narmour, 1991, p. 3)

Although the theory is presented in a music-analytic fashion, it has psychological relevance because it advances hypotheses about general perceptual principles which are precisely and quantitatively specified and therefore amenable to empirical investigation (Krumhansl, 1995b; Schellenberg, 1996).

In the bottom-up system, sequences of melodic intervals vary in the degree of *closure* that they convey according to the degree to which they exhibit the following characteristics:

1. an interval is followed by a rest;
2. the second tone of an interval has greater duration than the first;
3. the second tone occurs in a stronger metrical position than the first;
4. the second tone is more stable (less dissonant) in the established key or mode than the first;
5. three successive tones create a large interval followed by a smaller interval;
6. registral direction changes between the two intervals described by three successive tones.

Narmour (1990) provides rules for evaluating the influence of each condition on the closure conveyed by a sequence of intervals. While strong closure signifies the termination of ongoing melodic structure, an interval which is unclosed is said to be an *implicative interval* and generates expectations for the following interval which is termed the *realised interval*. The expectations generated by implicative intervals are described by Narmour (1990) in terms of several principles of continuation which are influenced by the Gestalt principles of proximity, similarity, and good continuation. In particular, according to the theory, small melodic intervals imply a *process* (the realised interval is in the same direction as the implicative interval and will be similar in size) and large melodic intervals imply a *reversal* (the realised interval is in a different direction to the implicative interval and is smaller in size).

The following description of the principles of the bottom-up system is based on an influential summary by Krumhansl (1995b). Some of these principles operate differently for small and large intervals which are defined by Narmour (1990) to be those of five semitones or less and seven semitones or more, respectively. The tritone is considered to be a threshold interval which may function as small or large (*i.e.,* implying continuation or reversal) depending on the context. The following principles make up the bottom-up system of melodic implication.

Registral direction states that small intervals imply continuations in the same registral direction whereas large intervals imply a change in registral direction (*cf.* the gap-fill process of Meyer, 1973). The application of the principle to small intervals is related to the Gestalt principle of good continuation.

Intervallic difference states that small intervals imply a subsequent interval that is similar in size (± 2 semitones if registral direction changes and ± 3 semitones if direction continues), while large intervals imply a consequent interval that is smaller in size (at least three semitones smaller if registral direction changes and at least four semitones smaller if direction continues). This principle can be taken as an application of the Gestalt principles of similarity and proximity for small and large intervals respectively.

Registral return is a general implication for a return to the pitch region (± 2 semitones) of the first tone of an implicative interval in cases where the realised interval reverses the registral direction of the implicative interval. Krumhansl (1995b) coded this principle as a dichotomy although Narmour (1990) distinguishes between *exact* and *near* registral return suggesting that the principle be graded as a function of the size of the interval between the realised tone and the first tone of the implicative interval (Schellenberg, 1996; Schellenberg *et al.*, 2002). This principle can be viewed as an application of the Gestalt principle of proximity in terms of pitch and similarity in terms of pitch interval.

Proximity describes a general implication for small intervals (five semitones or less) between any two tones. The implication is graded according to the absolute size of the interval. This principle can be viewed as an application of the Gestalt principle of proximity.

Closure is determined by two conditions: first, a change in registral direction;

> and second, movement to a smaller-sized interval. Degrees of closure exist corresponding to the satisfaction of both, one or neither of the conditions.[2]

In this encoding, the first three principles (registral direction, intervallic difference and registral return) assume dichotomous values while the final two (proximity and closure) are graded (Krumhansl, 1995b). Although the bottom-up IR principles are related to generic Gestalt principles, they are parameterised and quantified in a manner specific to music.

Narmour (1990) uses the principles of registral direction and intervallic difference to derive a complete set of 12 basic melodic structures each consisting of an implicative and a realised interval. These structures are described in Table 8.1 according to the direction of the realised interval relative to the implicative interval (same or different), the size of the realised interval relative to the implicative interval (larger, similar or smaller) and the size of the implicative interval (large or small). The resulting structures are classified into two groups: the *retrospective* structures are so-called because, although they differ in terms of the size of the implicative interval, they have the same basic shape and are heard in retrospect as variants of the corresponding *prospective structures*. While prospective interpretations of implications occur when the implied realisation actually occurs, retrospective interpretations occur when the implications are denied. The strength of the implications generated by each basic melodic structure depends on the degree to which it satisfies either or both of the principles of registral direction and intervallic difference. In an experimental study of the IR theory, Krumhansl (1995b) reports limited support for the basic melodic structures suggesting that expectations depend not only on registral direction and intervallic difference (which define the basic melodic structures) but also the principles of proximity, registral return and closure which are less explicitly formulated in the original presentation of the IR theory (Krumhansl, 1995b).

[2]Note that the principle of closure specifies the combinations of implicative and realised intervals that contribute to melodic closure (defined above) which signifies the termination of ongoing melodic structure and results in weak expectations.

Basic melodic structure	Implicative interval size	Direction of realised *cf.* implicative interval	Size of realised *cf.* implicative interval	Registral Direction	Intervallic Difference
Process, P	Small	Same	Similar	✓	✓
Intervallic Process, IP	Small	Different	Similar	✗	✓
Registral Process, VP	Small	Same	Larger	✓	✗
Retrospective Reversal, (R)	Small	Different	Smaller	✗	✗
Retrospective Intervallic Reversal, (IR)	Small	Same	Smaller	✓	✗
Retrospective Registral Reversal, (VR)	Small	Different	Larger	✗	✗
Reversal, R	Large	Same	Smaller	✓	✓
Intervallic Reversal, IR	Large	Different	Smaller	✗	✓
Registral Reversal, VR	Large	Same	Larger	✓	✗
Retrospective Process, (P)	Large	Different	Similar	✗	✗
Retrospective Intervallic Process, (IP)	Large	Same	Similar	✓	✗
Retrospective Registral Process, (VP)	Large	Different	Larger	✗	✗

Table 8.1: The basic melodic structures of the IR theory (Narmour, 1990).

In other respects, the quantitative model developed by Krumhansl (1995b) lacks some of the more complex components of the IR theory. For example, Narmour (1992) presents a detailed analysis of how the basic melodic structures combine together to form longer and more complex structural patterns of melodic implication within the IR theory. Three or more consecutive structures may form a *chain* in one of three ways depending on the closure implied by the antecedent structure: if there is sufficient closure, the antecedent structure will be *separated* from the subsequent structure (they share a tone); if closure is weak or suppressed, the structure will be *combined* with the subsequent structure (they share an interval); and finally, one structure may also be *embedded* in another. Chaining is encouraged by weak closure as measured by one or more of its contributing factors. Another way in which the IR theory addresses more complex melodic structure is through the emergence of higher hierarchical levels of structural representation when strong closure exists at lower levels. Structural tones (those beginning or ending a melodic structure, combination or chain) which are emphasised by strong closure at one level are said to *transform* to the higher level.

According to the theory, the same bottom-up principles of implication operate on sequences of (possibly non-contiguous) tones at higher transformational levels and, theoretically, a tone may be transformed to any number of higher levels. According to the theory, transformed tones may retain some of the registral implications of the lower level – an example of the primacy of the bottom-up aspects of the theory. Krumhansl (1997) has found some empirical support for the psychological validity of higher level implications in experiments with specially constructed melodic sequences. Finally, although quantitative implementations have tended to focus on the *parametric scales* of registral direction and interval size, the IR theory also includes detailed treatment of other parametric scales such as duration, metric emphasis and harmony (Narmour, 1990, 1992).[3]

The IR theory also stresses the importance of top-down influences on melodic expectancy. The top-down system is acquired on the basis of musical experience and, as a consequence, varies across musical cultures and traditions. The influences exerted by the top-down system include both *extra-opus* knowledge about style-specific norms, such as diatonic interpretations, tonal and metrical hierarchies and harmonic progressions, and *intra-opus* knowledge about aspects of a particular composition such as distinctive motivic and rhythmic patterns. Bharucha (1987) makes a similar distinction between *schematic* and *veridical*

[3]The status of these aspects of melody in the IR theory is criticised by some reviewers (Cross, 1995; Thompson, 1996).

influences on expectancy: while the former are influenced by schematic representations of typical musical relationships acquired through extensive exposure to a style, the latter are aroused by the activation of memory traces for specific pieces or prior knowledge for what is to come. Finally, the top-down system may generate implications that conflict with and potentially over-ride those generated by the bottom-up system. Efforts to develop quantitative implementations of the IR theory have tended to focus on the bottom-up system (see §8.2.3.2) with the top-down system represented only by relatively simple quantitative predictors (see §8.2.3.3).

It is important to emphasise that the present research is primarily concerned with those concrete implementations of the IR theory that, although they lack much of the music-analytic detail of Narmour's theory, have been examined in an empirical, psychological context. Although Narmour considered the five principles summarised above to be "a fair representation of his model" (Schellenberg, 1996, p. 77) and refers the reader to Krumhansl (1995b) amongst others for "evaluations of the model" (Narmour, 1999, p. 446), the present research is relevant to the IR theory of Narmour (1990, 1992) only to the extent that the concrete implementations examined are viewed as representative of the basic tenets of the theory. The IR theory has been the subject of several detailed reviews published in the psychological and musicological literature (Cross, 1995; Krumhansl, 1995b; Thompson, 1996) to which the reader is referred for more thorough summaries of its principal features.

8.2.3 Empirical Studies of Melodic Expectancy

8.2.3.1 Overview

Expectancy in music has been studied in experimental settings from a number of perspectives including the influence of rhythmic (Jones, 1987; Jones & Boltz, 1989), melodic (Cuddy & Lunny, 1995; Krumhansl, 1995b) and harmonic structure (Bharucha, 1987; Schmuckler, 1989). A variety of experimental paradigms have been employed to study expectancy including rating completions of musical contexts (Cuddy & Lunny, 1995; Krumhansl, 1995a; Schellenberg, 1996), generating continuations to musical contexts (Carlsen, 1981; Schmuckler, 1989; Thompson *et al.*, 1997; Unyk & Carlsen, 1987), classifying and remembering musical fragments (Schmuckler, 1997), reaction time experiments (Aarden, 2003; Bharucha & Stoeckig, 1986) and continuous response methodologies (Eerola *et al.*, 2002). Although expectancy in music has been shown to operate in a number of different contexts over a number of different

parameters and structural levels in music, this review is restricted to studies of expectancy in melodic music and, in particular, those which have specifically addressed the claims of the IR theory. Empirical research examining the bottom-up and top-down systems is discussed in §8.2.3.2 and §8.2.3.3 respectively.

8.2.3.2 The Bottom-up System

Cuddy & Lunny (1995) tested the bottom-up principles of the IR theory (as quantified by Krumhansl, 1995b) against goodness-of-fit ratings collected for continuation tones following a restricted set of two-tone melodic beginnings (see also §8.5). A series of multiple regression analyses supported the inclusion of intervallic difference, proximity and registral return in a theory of melodic expectancy. Support was also found for a revised version of registral direction which pertains to large intervals only and an additional bottom-up principle of pitch height, based on the observation that ratings tended to increase as the pitch height of the continuation tone increased. No support was found for the bottom-up principle of closure.

Krumhansl (1995a) repeated the study of Cuddy & Lunny (1995) with sixteen musically trained American subjects using a more complete set of two-tone contexts ranging from a descending major seventh to an ascending major seventh. Analysis of the results yielded support for modified versions of proximity, registral return and registral direction but not closure or intervallic difference. In particular, the results supported a modification of proximity such that it is linearly graded over the entire range of intervals used and a modification of registral return such that it varies as a linear function of the proximity of the third tone to the first.[4] Finally, the principle of registral direction was supported by the analysis except for the data for the major seventh which carried strong implications for octave completion (see also Carlsen, 1981). Support was also found for two extra principles that distinguish realised intervals forming octaves and unisons respectively. Krumhansl (1995a) also examined the effects of bottom-up psychophysical principles finding support for predictors coding the consonance of a tone with the first and second tones of the preceding interval (based on empirical and theoretical considerations).

Other experimental studies have extended these findings to expectations generated by exposure to melodic contexts from existing musical repertoires. Krumhansl (1995b) reports a series of three experiments: the first used eight melodic fragments taken from British folk songs, diatonic continuation tones

[4] As originally intended by Narmour (Schellenberg, 1996; Schellenberg *et al.*, 2002).

and twenty American subjects of whom 10 were musically trained and 10 untrained (see also §8.6); the second used eight extracts from Webern's *Lieder* (Opus 3, 4 and 15), chromatic continuation tones and 26 American subjects generally unfamiliar with the atonal style of whom 13 were musically trained and 13 untrained; and the third used 12 melodic fragments from Chinese folk songs, pentatonic continuation tones and 16 subjects of whom 8 were Chinese and 8 American. All the melodic contexts ended on an implicative interval and all continuation tones were within a two octave range centred on the final tone of the context. Analysis of the results yielded support for all of the bottom-up principles (with the exception of intervallic difference for the second experiment). Overall, the weakest contribution was made by intervallic difference and the strongest by proximity. Support was also found for the unison principle of Krumhansl (1995a).

Schellenberg (1996) argued that the bottom-up models discussed above are overspecified and contain redundancy due to collinearities between their component principles. As a result, the theory may be expressed more simply and parsimoniously without loss of predictive power. Support was found for this argument in an independent analysis of the experimental data first reported by Krumhansl (1995b) using a model consisting of registral return, registral direction revised such that it applies only to large intervals (although quantified in a different manner to the revision made by Cuddy & Lunny, 1995) and a revised version of proximity (similar in spirit, though quantitatively different, to the revision made by Krumhansl, 1995a). In a further experiment, Schellenberg (1997) applied principal components analysis to this revised model which resulted in the development of a two-factor model. The first factor is the principle of proximity as revised by Schellenberg (1996); the second, *pitch reversal* is an additive combination of the principles of registral direction (revised) and registral return. This model is considerably simpler and more parsimonious than Schellenberg's revised model and yet does not compromise the predictive power of that model in accounting for the data obtained by Krumhansl (1995b) and Cuddy & Lunny (1995).

Similar experiments with Finnish spiritual folk hymns (Krumhansl *et al.*, 1999) and indigenous folk melodies (yoiks) of the Sami people of Scandinavia (Krumhansl *et al.*, 2000) have, however, questioned the cross-cultural validity of such revised models. In both studies, it was found that the model developed by Krumhansl (1995a) provided a much better fit to the data than those of Krumhansl (1995b) and Schellenberg (1996, 1997). By contrast, Schellenberg *et al.* (2002) have found the opposite to be true in experiments with adults and

infants in a task involving the rating of continuation tones following contexts taken from Acadian (French Canadian) folk songs. They suggest that the difference may be attributable partly to the fact that none of the musical contexts used in the experiments of Krumhansl *et al.* (1999, 2000) ended in unambiguously large and implicative intervals (Schellenberg *et al.*, 2002, p. 530). While Schellenberg *et al.* (2002) and Krumhansl *et al.* (1999) found strong support for the principle of proximity with only limited influence of registral return and intervallic difference, Krumhansl *et al.* (2000) found the strongest bottom-up influence came from the principle of intervallic difference with weak support for the principles of proximity and registral return. The consonance predictors of Krumhansl (1995a) made a strong contribution to both models especially in the case of the folk hymns (Krumhansl *et al.*, 1999, 2000).

According to the IR theory, the principles of the bottom-up system exert a consistent influence on expectations regardless of the musical experience of the listener and the stylistic context notwithstanding the fact that the expectations actually generated are predicted to be subject to these top-down influences. Indirect support for this claim comes in the form of high correlations between the responses of musically trained and untrained subjects (Cuddy & Lunny, 1995; Schellenberg, 1996) and between the responses of groups with different degrees of familiarity with the musical style (Eerola, 2004a; Krumhansl *et al.*, 2000; Schellenberg, 1996). Regardless of the cognitive mechanisms on which they depend, melodic expectations tend to exhibit a high degree of similarity across levels of musical training and familiarity. Further evidence is provided by qualitatively similar degrees of influence of the bottom-up principles on the expectations of musically trained and untrained subjects (Cuddy & Lunny, 1995; Schellenberg, 1996) and across levels of relevant stylistic experience (Krumhansl *et al.*, 1999; Schellenberg, 1996). These findings have typically been interpreted as support for the universality of the bottom-up principles.

However, there are several reasons to question this conclusion. First, other research on melodic expectancy has uncovered differences across levels of training. von Hippel (2002), for example, conducted an experiment in which trained and untrained subjects were asked to make prospective contour judgements for a set of artificially generated melodies. While the expectations of the trained listeners exhibited the influence of pitch reversal and *step momentum* (the expectation that a melody will maintain its registral direction after small intervals) the responses of the untrained listeners exhibited significantly weaker influences of these principles. Furthermore, in a study of goodness-of-fit rat-

ings of single intervals as melodic openings and closures, Vos & Pasveer (2002) found that the responses of untrained listeners exhibited a greater influence of intervallic direction than those of the trained listeners.

Second, it must be noted that the empirical data cover a limited set of cultural groups and differences in observed patterns of expectation related to cultural background have been found (Carlsen, 1981). Furthermore, some studies have uncovered cross-cultural differences in the strength of influence of the bottom-up principles on expectancy. Krumhansl *et al.* (2000), for example, found that the correlations of the predictors for intervallic difference, registral return and proximity were considerably stronger for the Western listeners than for the Sami and Finnish listeners. Eerola (2004a) made similar observations in a replication of this study with traditional healers from South Africa.

Third, the influence of the bottom-up principles appears to vary with the musical stimuli used. Krumhansl *et al.* (2000) note that while the Finnish listeners in their study of expectancy in Sami folk songs exhibited a strong influence of consonance, the Finnish listeners in the earlier study of expectancy in Finnish hymns (Krumhansl *et al.*, 1999) exhibited a weaker influence of consonance in spite of having a similar musical background. Krumhansl *et al.* (2000) suggest that this may indicate that the Finnish listeners in their study adapted their judgements to the relatively large number of consonant intervals present in their experimental materials. More generally, the research reviewed in this section diverges significantly in the support found for the original bottom-up principles, revised versions of these principles and new principles. The most salient differences between the studies, and the most obvious cause of this discrepancy, are the musical contexts used to elicit expectations. Krumhansl *et al.* (2000, p. 41) conclude that "musical styles may share a core of basic principles, but that their relative importance varies across styles."

The influence of melodic context on expectations has been further studied by Eerola *et al.* (2002) who used a continuous response methodology (see 8.7.1) to collect subjects' continuous judgements of the predictability of melodies (folk songs, songs composed by Charles Ives and isochronous artificially generated melodies) simultaneously as they listened to them. The predictability ratings were analysed using three models: first, the IR model; second, a model based on the entropy (see §6.2.2) of a monogram distribution of pitch intervals with an exponential decay within a local sliding window (the initial distribution was derived from an analysis of the EFSC, see Chapter 4); and third, a variant of the second model in which the pitch class distribution was used and was initialised using the key profiles of Krumhansl & Kessler (1982). The re-

sults demonstrated that the second model and, in particular, the third model accounted for much larger proportions of the variance in the predictability data than the IR model while a linear combination of the second and third models improved the fit even further (Eerola, 2004b). It was argued that the success of these models was a result of their ability to account for the data-driven influences of melodic context.

Finally, it is important to note that universality or ubiquity of patterns of behaviour does not imply innateness. To the extent that the bottom-up principles capture universal patterns of behaviour, they may reflect the influence of long-term informal exposure to simple and ubiquitous regularities in music (Schellenberg, 1996; Thompson *et al.*, 1997). In accordance with this position, Bergeson (1999) found that while adults are better able to detect a pitch change in a melody that fulfils expectations according to the IR theory (Narmour, 1990) than in one that does not, six and seven month old infants do not exhibit this difference in performance across conditions. In addition, Schellenberg *et al.* (2002) report experiments examining melodic expectancy in adults and infants (covering a range of ages) using experimental tasks involving both rating and singing continuation tones to supplied melodic contexts. The data were analysed in the context of the IR theory as originally formulated (Schellenberg, 1996) and as revised by Schellenberg (1997). The results demonstrate that expectations were better explained by both models with increasing age and musical exposure. While consecutive pitch proximity (Schellenberg, 1997) was a strong influence for all listeners, the influence of more complex predictors such as pitch reversal (Schellenberg, 1997) and registral return (Schellenberg, 1996) only became apparent with the older listeners. Schellenberg *et al.* (2002) conclude with a discussion of possible explanations for the observed developmental changes in melodic expectancy: first, they may reflect differences between infant-directed speech and adult-directed speech; second, they may reflect general developmental progressions in perception and cognition (*e.g.,* perceptual differentiation and working or sensory memory), which exert influence across domains and modalities; and third, they may reflect increasing exposure to music and progressive induction of increasingly complex regularities in that music.

8.2.3.3 The Top-down System

In addition to studying the bottom-up principles of the IR theory, research has also examined some putative top-down influences on melodic expectation many of which are based on the key profiles of perceived tonal stability empirically

quantified by Krumhansl & Kessler (1982). Schellenberg (1996) and Krumhansl (1995b), for example, found support for the inclusion in a theory of expectancy of a tonality predictor based on the key profile for the major or minor key of the melodic fragment (see §3.6). Cuddy & Lunny (1995) examined the effects of several top-down tonality predictors. The first consisted of four tonal hierarchy predictors similar to those of Schellenberg (1996) and Krumhansl (1995b) based on the major and minor key profiles for the first and second notes of the context interval. The second, *tonal strength*, was based on the assumption that the rating of a continuation tone would be influenced by the degree to which the pattern of three tones suggested a tonality.[5] The third tonality predictor, *tonal region*, was derived by listing all possible major and minor keys in which each implicative interval was diatonic and coding each continuation tone according to whether it represented a tonic of one of these keys. Support was found for all of these top-down influences although it was also found that the predictors for tonal hierarchy could be replaced by tonal strength and tonal region without loss of predictive power. Krumhansl (1995a) extended the tonal region predictor developed by Cuddy & Lunny (1995) by averaging the key profile data for all keys in which the two context tones are diatonic. Strong support was found for the resulting predictor variable for all context intervals except for the two (ascending and descending) tritones. In contrast, no support was found for the tonal strength predictor of Cuddy & Lunny (1995).

While neither Cuddy & Lunny (1995) nor Schellenberg (1996) found any effect of musical training on the influence of top-down tonality predictors, Vos & Pasveer (2002) found that the consonance of an interval (based on music-theoretical considerations) influenced the goodness-of-fit judgements of the trained listeners to a much greater extent than those of the untrained listeners in their study of intervals as candidates for melodic openings and closures. In a further analysis of their own data, Krumhansl *et al.* (1999) sought to distinguish between schematic and veridical top-down influences on expectations (Bharucha, 1987, see §8.2.2). The schematic predictors were the two-tone continuation ratings obtained by Krumhansl (1995a) and the major and minor key profiles (Krumhansl & Kessler, 1982). The veridical predictors consisted of monogram, digram and trigram distributions of tones in the entire corpus of spiritual folk hymns and a predictor based on the correct continuation tone. It was found that the schematic predictors showed significantly stronger effects for the non-experts in the study than the experts. In contrast, veridical predictors such as monogram and trigram distributions and the correct next tone

[5] The key-finding algorithm developed by Krumhansl and Schmuckler (Krumhansl, 1990) was used to rate each of the patterns for tonal strength.

showed significantly stronger effects for the experts than for the non-experts. Krumhansl *et al.* (2000) found similar effects in their study of North Sami yoiks and showed that these effects were related to familiarity with individual pieces used in the experiment. These findings suggest that increasing familiarity with a given stylistic tradition tends to weaken the relative influence of top-down schematic knowledge of Western tonal-harmonic music on expectancy and increase the relative influence of specific veridical knowledge of the style.

There is some evidence, however, that the rating of continuation tones may elicit schematic tonal expectations specifically related to melodic closure since the melody is paused to allow the listener to respond. Aarden (2003) reports an experiment in which subjects were asked to make retrospective contour judgements for each event in a set of European folk melodies. Reaction times were measured as an indication of the strength and specificity of expectations under the hypothesis that strong and accurate expectations facilitate faster responses (see also Bharucha & Stoeckig, 1986). The resulting data were analysed using the two-factor model of Schellenberg (1997). While a tonality predictor based on the key profiles of Krumhansl & Kessler (1982) made no significant contribution to the model, a monogram model of pitch frequency in the EFSC (see Chapter 4) did prove to be a significant predictor. In a second experiment, subjects were presented with a counter indicating the number of notes remaining in the melody and asked to respond only to the final tone. In this case, the Krumhansl & Kessler tonality predictor, which bears more resemblance to the distribution of phrase-final tones than that of all melodic tones in the EFSC, made a significant contribution to the model. On the basis of these results, Aarden (2003) argues that the schematic effects of tonality may be limited to phrase endings whereas data-driven factors, directly reflecting the structure and distribution of tones in the music, have more influence in melodic contexts that do not imply closure.

Finally, it is worth noting that the top-down tonality predictors that have been examined in the context of modelling expectation have typically been rather simple. In this regard, Povel & Jansen (2002) report experimental evidence that goodness ratings of entire melodies depend not so much on the overall stability of the component tones (Krumhansl & Kessler, 1982) but the ease with which the listener is able to form a harmonic interpretation of the melody in terms of both the global harmonic context (key and mode) and the local movement of harmonic regions. The latter process is compromised by the presence of non-chord tones to the extent that they cannot be assimilated by means of anchoring (Bharucha, 1984) or by being conceived as part of a run

of melodic steps. Povel & Jansen (2002) argue that the harmonic function of a region determines the stability of tones within that region and sets up expectations for the resolution of unstable tones.

8.2.3.4 Summary

While the results of many of the individual studies reviewed in the foregoing sections have been interpreted in favour of the IR theory, the overall pattern emerging from this body of research is rather different. Empirical research has demonstrated that some collection of principles based on the bottom-up IR system can generally account rather well for the patterns of expectation observed in a given experiment but it is also apparent that any such set constitutes too inflexible a model to fully account for the effects of differences across experimental settings in terms of the musical experience of the listeners and the melodic contexts in which expectations are elicited. Regarding the top-down system, research suggests that the expectations of listeners show strong effects of schematic factors such as tonality although the predictors typically used to model these effects may be too inflexible to account for the effects of changing the context in which expectations are elicited.

8.3 Statistical Learning of Melodic Expectancy

8.3.1 The Theory

A theory of the cognitive mechanisms underlying the generation of melodic expectations is presented here. It is argued that this theory is capable of accounting more parsimoniously for the behavioural data than the quantitative formulations of the IR theory while making fewer assumptions about the cognitive mechanisms underlying the perception of music. From the current perspective, the quantitatively formulated principles of the IR theory provide a descriptive, but not explanatory, account of expectancy in melody: they describe human behaviour at a general level but do not account for the cognitive mechanisms underlying that behaviour. To the extent that the two theories produce similar predictions, they are viewed as lying on different levels of explanation (Marr, 1982; McClamrock, 1991). Both bottom-up and top-down components of the quantitatively formulated IR models have been found to provide an inadequate account of the detailed influences of musical experience and musical context on melodic expectancy (see §8.2.3). The theory proposed here is motivated by the need to formulate a more comprehensive account of these influences.

In particular, the present theory questions the need, and indeed the validity, of positing a distinction between bottom-up and top-down influences on expectation, and especially the claim that the principles of the bottom-up system reflect innately specified representations of sequential dependencies between musical events. According to the theory, the bottom-up principles of the IR theory constitute a description of common regularities in music which are acquired as mature patterns of expectation through extensive exposure to music. Rather than invoking innate representational rules (such as the bottom-up principles and the basic melodic structures of the IR theory), this theory invokes innate general purpose learning mechanisms which impose architectural rather than representational constraints on cognitive development (Elman *et al.*, 1996). Given exposure to appropriate musical stimuli, these learning mechanisms can acquire domain specific representations and behaviour which is approximated by the principles of the IR theory (see also Bharucha, 1987; Gjerdingen, 1999b).

It is hypothesised that the bottom-up principles of the quantitatively formulated IR models (as well as other proposed bottom-up influences on expectancy) reflect relatively simple musical regularities which display a degree of pan-stylistic ubiquity. To the extent that this is the case, these bottom-up IR principles are regarded as formalised approximate descriptions of the mature behaviour of a cognitive system that acquires representations of the statistical structure of the musical environment. On the other hand, top-down factors, such as tonality, reflect the induction of rather more complex musical structures which show a greater degree of variability between musical styles. If this is indeed the case, a single learning mechanism may be able to account for the descriptive adequacy of some of the bottom-up principles across degrees of expertise and familiarity as well as for differences in the influence of other bottom-up principles and top-down factors. By replacing a small number of symbolic rules with a general-purpose learning mechanism, the theory can account more parsimoniously for both consistent and inconsistent patterns of expectation between groups of listeners on the basis of differences in prior musical exposure, the present musical context and the relative robustness of musical regularities across stylistic traditions.

8.3.2 Supporting Evidence

We shall discuss existing evidence that supports the proposed theory of expectancy in terms of the necessary conditions that must be satisfied for the theory to hold. In particular we ask two questions: Are the regularities in music sufficient to support the acquisition of the experimentally observed patterns

of melodic expectation? And: Is there any evidence that listeners possess cognitive mechanisms that are capable of acquiring such behaviour through exposure to music?

Regarding the first question, research suggests that expectancy operates very similarly in tasks which elicit ratings of continuations to supplied melodic contexts (see §8.2.3) and tasks which elicit spontaneous production of continuations to melodic contexts (Schellenberg, 1996; Schmuckler, 1989, 1990; Thompson *et al.*, 1997). If the perception and production of melodies are influenced by similar principles, it is pertinent to ask whether existing repertoires of compositions also reflect such influences of melodic implication. Thompson & Stainton (1996, 1998) have examined the extent to which the bottom-up principles of the IR theory are satisfied in existing musical repertoires including the soprano and bass voices of chorales harmonised by J. S. Bach, melodies composed by Schubert and Bohemian folk melodies. Preliminary analyses indicated that significant proportions of implicative intervals satisfy the principles of intervallic difference, registral return and proximity while smaller proportions satisfied the other bottom-up principles. The proportions were highly consistent across the three datasets. Furthermore, a model consisting of the five bottom-up principles accounted for much of the variance in the pitch of tones following implicative intervals in the datasets (as well as closural intervals in the Bohemian folk melodies – Thompson & Stainton, 1998). With the exception of intervallic difference for the Schubert dataset, all five principles contributed significantly to the predictive power of the model. These analyses demonstrate that existing corpora of melodic music contain regularities that tend to follow the predictions of the IR theory and that are, in principle, capable of supporting the acquisition of patterns of expectation that accord with its principles.

Given these findings, an argument can be made that the observed regularities in music embodied by the bottom-up IR principles reflect universal physical constraints of performance rather than attempts to satisfy universal properties of the perceptual system. Examples of such constraints include the relative difficulty of singing large intervals accurately and the fact that large intervals will tend towards the limits of a singer's vocal range (Russo & Cuddy, 1999; Schellenberg, 1997). von Hippel & Huron (2000) report a range of experimental evidence supporting the latter observation as an explanation of *post-skip reversals* (*cf.* the gap-fill pattern of Meyer, 1973 and the principles of registral direction and registral return of Narmour, 1990) which they account for in terms of *regression towards the mean* necessitated by tessitura. In one experiment, for example, it was found that evidence for the existence of post-skip reversals in

a range of musical styles is limited to those skips (intervals of three semitones or more) that cross or move away from the median pitch of a given corpus of music. When skips approach the median pitch or land on it, there is no significant difference in the proportions of continuations and reversals of registral direction. In spite of this, von Hippel (2002) found that the expectations of listeners actually reflect the influence of perceived post-skip reversals suggesting that patterns of expectation are acquired as heuristics representing simplified forms of more complex regularities in music.

We turn now to the question of whether the cognitive mechanisms exist to acquire the observed patterns of melodic expectation through exposure to existing music. Saffran *et al.* (1999) have elegantly demonstrated that both adults and eight month old infants are capable of learning to segment continuous tone sequences on the basis of differential transitional probability distributions of tones within and between segments. On the basis of these and similar results with syllable sequences, Saffran *et al.* (1999) argue that human infants and adults possess domain general learning mechanisms which readily compute transitional probabilities on exposure to auditory sequences. Furthermore, Oram & Cuddy (1995) conducted a series of experiments in which continuation tones were rated for musical fit in the context of artificially constructed sequences of pure tones in which the tone frequencies were carefully controlled. The continuation tone ratings of both trained and untrained listeners were significantly related to the frequency of occurrence of the continuation tone in the context sequence. Cross-cultural research has also demonstrated the influence of tone distributions on the perception of music (Castellano *et al.*, 1984; Kessler *et al.*, 1984; Krumhansl *et al.*, 1999). In particular, Krumhansl *et al.* (1999) found significant influences of second order distributions on the expectations of the expert listeners in their study (see §8.2.3.2).

There is also evidence that listeners are sensitive to statistical regularities in the size and direction of pitch intervals in the music they are exposed to. In a statistical analysis of a large variety of Western melodic music, for example, Vos & Troost (1989) found that smaller intervals tend to be of a predominantly descending form while larger ones occur mainly in ascending form. A behavioural experiment demonstrated that listeners are able to correctly classify artificially generated patterns that either exhibited or failed to exhibit the regularity. Vos & Troost consider two explanations for this result: first, that it is connected with the possibly universal evocation of musical tension by ascending large intervals and of relaxation by descending small intervals (Meyer, 1973); and second, that it reflects overlearning of conventional musical patterns. Vos & Troost do

not strongly favour either account, each of which depends on the experimentally observed sensitivity of listeners to statistical regularities in the size and direction of melodic intervals.

In summary, research has demonstrated that existing repertoires of music exhibit regularities which broadly follow the predictions of the bottom-up principles of the IR theory and which, in some cases, may be related to physical constraints of performance. Furthermore, there is evidence that listeners are sensitive to statistical regularities in music and that these regularities are exploited in the perception of music.

8.3.3 The Model

The theory of melodic expectancy presented in §8.3.1 predicts that it should be possible to design a statistical learning algorithm, such as the one developed in Chapters 6 and 7, with no initial knowledge of sequential dependencies between melodic events which, given exposure to a reasonable corpus of music, would exhibit similar patterns of melodic expectation to those observed in experiments with human subjects (see also Bharucha, 1993).

The computational system developed in Chapters 6 and 7 provides an attractive model of melodic expectancy for several reasons. First, in accordance with the theory put forward here, while it is endowed (via the multiple viewpoints framework) with sensitivities to certain musical features, the untrained system has no structured expectations about sequential melodic patterns. Any structure found in the patterns of expectation exhibited by the trained system is a result of statistical induction of regularities in the training corpus. Second, the system simulates at a general level of description the situation of the listener attending to a melody. When exposed to a novel melody, the model exhibits patterns of expectation about forthcoming events based on the foregoing melodic context as do human listeners. Furthermore, these patterns of expectation are influenced both by existing extra-opus and incrementally increasing intra-opus knowledge (see §6.2.4).

Finally, the patterns of expectation exhibited by the system are sensitive to many different musical attributes which are motivated by previous research on music perception as discussed in §5.4. In this regard, note that both the present theory and the IR theory assume a sensitivity to certain features of discrete events making up the musical surface. If the system is to account for the same experimental observations as the two-factor model (Schellenberg, 1997), it must be capable of representation and inference over the dimensions of size and direction of melodic intervals (in which terms the bottom-up principles

are expressed) and pitches and scale degrees (in which terms the top-down principles have generally been expressed).

It is important to make a clear distinction between the cognitive theory of melodic expectancy presented in §8.3.1 and the statistical model of melodic expectancy embodied in the computational system presented in Chapters 6 and 7. In the remainder of the chapter, unless otherwise specified, references to *the theory* or *the model* will honour this distinction (see §2.6). While the computational model embodies the theory, to the extent that it relies purely on statistical learning, it also goes well beyond the theory in the details of its implementation. This is necessary in order for it to exhibit behaviours of any complexity (see also §8.8).

8.4 Experimental Methodology

The present research has two primary objectives which, in accordance with the level at which the theory is presented (and the manner in which it diverges from the IR theory), are stated at a rather high level of description. The first objective is to test the hypothesis that the statistical model developed in Chapters 6 and 7 is able to account for the patterns of melodic expectation observed in experiments with human subjects at least as well as the IR theory. Since the statistical model acquires its knowledge of sequential melodic structure through exposure to melodic music, corroboration of the hypothesis would demonstrate that it is not necessary to posit innate and universal musical rules to account for the observed patterns of melodic expectation; melodic expectancy can be accounted for in terms of statistical induction of both intra- and extra-opus regularities in existing musical corpora.

The methodological approach followed in examining this hypothesis compares the patterns of melodic expectation generated by the computational model to those of human listeners observed in previously reported experiments. Three experiments are presented which elicit expectations in increasingly complex melodic contexts: first, in the context of the single intervals used by Cuddy & Lunny (1995); second, in the context of the excerpts from British folk songs used by Schellenberg (1996); and third, throughout the two chorale melodies used by Manzara *et al.* (1992).

In each experiment, the statistical models are compared with the two-factor model of Schellenberg (1997) plus a tonality predictor. Although the two-factor model did not perform as well as that of Krumhansl (1995a) in accounting for the expectations of the listeners in the experiments of Krumhansl *et al.*

(1999, 2000), the converse was true in the experiments of Schellenberg *et al.* (2002). While the debate surrounding the precise quantitative formulation of the bottom-up system appears likely to continue, this particular IR variant was chosen from those reviewed in §8.2.3.2 because it provides the most parsimonious formulation of the bottom-up principles without loss of predictive power in accounting for the data collected by Cuddy & Lunny (1995) and Schellenberg (1996) which are used in Experiments 1 and 2 respectively. Following common practice, the two-factor model was supplemented with a tonality predictor developed in previous research. In the first experiment, the influence of tonality was modelled using the tonal region predictor of Krumhansl (1995a) while the second and third experiments used the Krumhansl & Kessler key profiles for the notated key of the context.

Following Cutting *et al.* (1992) and Schellenberg *et al.* (2002), the statistical model and the two-factor model of expectancy are compared on the basis of scope, selectivity and simplicity. Regarding the scope of the two models, since the individual subject data were not available for any of the experiments and the two models are not nested, Williams' t statistic for comparing dependent correlations (Hittner *et al.*, 2003; Steiger, 1980) was used to compare the two models in each experiment. It is expected that the relative performance of the statistical model will increase with longer and more realistic melodic contexts. The selectivity of the models was assessed by using each model to predict random patterns of expectation in the context of the experimental stimuli used in each experiment. Finally, with regard to simplicity, we examine the extent to which the statistical model subsumes the function of the bottom-up components of the two-factor model in accounting for the behavioural data used in each experiment. An alpha level of 0.05 is used for all statistical tests.

The second objective is to examine which musical attributes present in, or simply derivable from, the musical surface afford regularities that are capable of supporting the acquisition of the empirically observed patterns of melodic expectation. In each experiment, hypotheses are presented regarding the specific attributes likely to afford such regularities. The approach taken to testing these hypotheses has been to select sets of viewpoints which maximise the fit between experimentally determined human patterns of expectation and those exhibited by the computational model. The selection of viewpoints was achieved using the forward stepwise selection algorithm described in §7.4. The use of forward selection and the preference for feature deletions over additions may be justified by the observation that simplicity appears to be a powerful and general organising principle in perception and cognition (Chater, 1999; Chater &

Vitányi, 2003). The performance measure used in these experiments is the regression coefficient of the observed human patterns of expectation (*e.g.,* mean continuation tone ratings) for a given set of melodic stimuli regressed on the patterns of expectation exhibited by the statistical model (using a given set of features) for the same set of stimuli. The evaluation functions used will be described in more detail for each experiment in turn (see §8.5, §8.6 and §8.7).

The feature sets used in these experiments consist of subsets of the attribute types shown in Tables 5.2 and 5.4 which were described and motivated in terms of previous research on music perception and cognition in §5.4. The corpus used to train the models consisted of Datasets 1, 2 and 9 (see Chapter 4). In discussing the experimental results, we shall talk about finding support for the influence of a particular feature on melodic expectancy. It should be kept in mind that this shorthand is intended to convey that support has been found for the existence of statistical regularities in a given melodic dimension that increase the fit between the behaviour of the model and the observed human behaviour.

8.5 Experiment 1

8.5.1 Method

The objective in this experiment was to examine how well the statistical model accounts for patterns of expectation following single interval contexts. Cuddy & Lunny (1995) report an experiment in which listeners were asked to rate continuation tones following a two tone context. The subjects were 24 undergraduate students at Queen's University in Canada of whom half were musically trained and half untrained. The stimuli consisted of eight implicative contexts corresponding to ascending and descending intervals of a major second, a minor third, a major sixth and a minor seventh. All subjects heard half of the contexts ending on C_4 and half ending on $F\sharp_4$ (see Table 8.2) in an attempt to discourage them from developing an overall top-down sense of tonality for the entire experiment. Continuation tones consisted of all 25 chromatic tones from one octave below to one octave above the second tone of the implicative context. The two tones of each context were presented as a dotted minim followed by a crotchet while all continuation tones had a minim duration. These durations were chosen to create a sense of 4/4 metre continuing from the first bar (containing the implicative interval) to the second bar (containing the continuation tone).

The subjects were asked to rate how well the continuation tone continued

Context interval		Second tone	
Interval	Direction	C	F♯
Major second	Ascending	B♭$_3$–C$_4$	E$_4$–F♯$_4$
	Descending	D$_4$–C$_4$	G♯$_4$–F♯$_4$
Minor third	Ascending	A$_3$–C$_4$	D♯$_4$–F♯$_4$
	Descending	E♭$_4$–C$_4$	A$_4$–F♯$_4$
Major sixth	Ascending	E♭$_3$–C$_4$	A$_3$–F♯$_4$
	Descending	A$_4$–C$_4$	D♯$_5$–F♯$_4$
Minor seventh	Ascending	D$_3$–C$_4$	G♯$_3$–F♯$_4$
	Descending	B♭$_4$–C$_4$	E$_5$–F♯$_4$

Table 8.2: The melodic contexts used in Experiment 1 (after Cuddy & Lunny, 1995, Table 2).

the melody on a scale from 1 (extremely bad continuation) to 7 (extremely good continuation). The experiment yielded 200 continuation tone ratings for each subject. An analysis of variance with the factors musical training, context interval and continuation tone yielded one significant interaction between context interval and continuation tone. Since there was no effect of training and the data exhibited high inter-subject correlation, the ratings for each continuation tone were averaged across subjects and training levels. The mean continuation tone ratings for trained and untrained subjects are available in Cuddy & Lunny (1995, Appendix).

In the present experiment, the trained model was exposed to each of the eight contexts used by Cuddy & Lunny (1995) for all of which the second tone was F♯$_4$. Due to the short contexts involved, the short-term model was not used in this experiment. In each case, the model returns a probability distribution over the set of 25 chromatic pitches ranging from F♯$_3$ to F♯$_5$. Since the distributions returned by the model are constrained to sum to one and are likely to violate the parametric normality assumption, each of the pitches was assigned a rank according to its estimated probability in inverse order (such that high probability pitches were assigned high ranks). The regression coefficient of the mean ratings obtained by Cuddy & Lunny (1995) regressed on the distribution ranks of the model was used as a performance metric in viewpoint selection. In terms of features used, chromatic pitch (`cpitch`) and pitch class (`cpitch-class`, see Shepard, 1982) were included although they were not expected to exert significant influences on expectancy as a result of the limited context. It was hypothesised that more abstract melodic features such as chromatic pitch interval (`cpint`) and interval class (`cpcint`) would be the

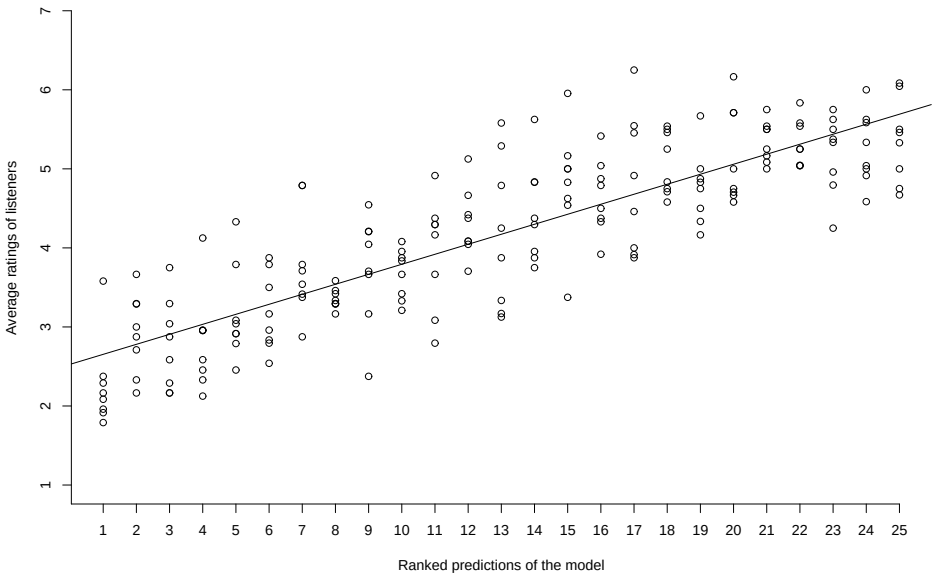


Figure 8.1: Correlation between subjects' mean goodness-of-fit ratings and the predictions of the statistical model for continuation tones in the experiments of Cuddy & Lunny (1995).

most important source of regularities underlying melodic expectancy (Dowling & Bartlett, 1981). Pitch contour (`contour`) was also included to examine the effects of a still more abstract representation of registral direction (Dowling, 1994). It was also hypothesised that the patterns of expectation may reflect a mode of perception in which subsequent tones are appraised in relation to the first tone in the context (`cpintfip`). Given the impoverished context, a sense of tonality may have been inferred based on the first tone of the context as tonic (Cohen, 2000; Cuddy & Lunny, 1995; Thompson *et al.*, 1997). In spite of the limited context, it was also hypothesised that pitch may have interacted with rhythmic dimensions of the contexts to generate expectations (Jones, 1987; Jones & Boltz, 1989). Consequently, a set of linked viewpoints (see §5.4) was included in the experiment which modelled interactions between three simple pitch-based attributes (`cpitch`, `cpint` and `contour`) and three rhythmic attributes (`dur`, `dur-ratio` and `ioi`).

8.5.2 Results

The final multiple viewpoint system selected in this experiment enabled the statistical model to account for approximately 72% of the variance in the mean continuation tone ratings [$R = 0.846, R^2_{adj} = 0.715, F(1, 198) = 500.2, p <$

Stage	Viewpoint Added	R
1	`cpint`⊗`dur`	0.775
2	`cpintfip`	0.840
3	`cpcint`	0.846

Table 8.3: The results of viewpoint selection in Experiment 1.

$0.001]$. The relationship between the patterns of expectation exhibited by the model and by the subjects in the experiments of Cuddy & Lunny (1995) is plotted with the fitted regression line in Figure 8.1. The statistical model provided a slightly closer fit to the data than the two-factor model, which accounted for approximately 68% of the variance in the data $[R = 0.827, R^2_{adj} = 0.679, F(3, 196) = 141.2, p < 0.001]$, although the difference was found not to be significant $[t(197) = 1.102, p = 0.272]$.

In order to examine the hypothesis that the statistical model subsumes the function of the bottom-up components of the two-factor model, a more detailed comparison of the two models was conducted. The expectations of the statistical model exhibit significant correlations in the expected directions with both components of the two-factor model: Proximity $[r(198) = -0.670,\ p < 0.001]$; and Reversal, $[r(198) = 0.311,\ p < 0.001]$. Furthermore, the fit of the statistical model to the behavioural data was not significantly improved by adding Proximity $[F(1, 197) = 1.537,\ p = 0.217]$, Reversal $[F(1, 197) = 0.001,\ p = 0.975]$ or both of these factors $[F(2, 196) = 0.809,\ p = 0.45]$ to the regression model. This analysis indicates that the statistical model entirely subsumes the function of Proximity and Reversal in accounting for the data collected by Cuddy & Lunny (1995).

Finally, in order to examine the selectivity of the two models, 50 sets of ratings for the stimuli ($N = 200$ in each set) were generated through random sampling from a normal distribution with a mean and SD equivalent to those of the listeners' ratings. With an alpha level of 0.05, just two of the 50 random vectors were fitted at a statistically significant level by each of the models and there was no significant difference between the fit of the two models for any of the 50 trials. Neither model is broad enough in its scope to successfully account for random data.

The results of viewpoint selection are shown in Table 8.3. As predicted on the basis of the short contexts, the viewpoints selected tended to be based on pitch interval structure. The limited context for the stimulation of expectancy is probably insufficient for the evocation of statistical regularities in chromatic pitch structure. The fact that `cpint`⊗`dur` was selected over and above its

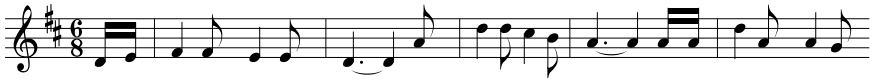
primitive counterpart (`cpint`) suggests that expectations were influenced by the interaction of regularities in pitch interval and duration. It might appear surprising that regularities in rhythmic structure should influence expectations with contexts so short. Although this may be an artefact, recall that Cuddy & Lunny (1995) carefully designed the rhythmic structure of their stimuli to induce a particular metric interpretation. The issue could be investigated further by systematically varying the rhythmic structure of the stimuli used to obtain goodness-of-fit ratings. Finally, the results reveal a strong influence of `cpintfip` on expectancy which may be partly accounted for by the brevity of the contexts, which do not contain enough information to reliably induce a tonality, combined with the relatively long duration of the first tone. Regularities in the three selected dimensions of existing melodies are such that the statistical model provides an equally close fit to the patterns of expectation observed in the experiment of Cuddy & Lunny (1995) as the two-factor model.

8.6 Experiment 2

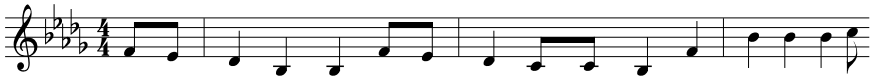
8.6.1 Method

The objective of this experiment was to extend the approach of Experiment 1 to patterns of expectation observed after longer melodic contexts drawn from an existing musical repertoire. Schellenberg (1996, Experiment 1) reports an experiment in which listeners were asked to rate continuation tones following eight melodic fragments taken from British folk songs (Palmer, 1983; Sharp, 1920). The subjects were 20 members of the community of Cornell University in the USA of whom half had limited musical training and half had moderate musical training. Figure 8.2 shows the eight melodic contexts of which four are in a minor mode and four in a major mode. They were chosen such that they ended on an implicative interval (see §8.2.2). Four of the fragments end with one of two small intervals (2 or 3 semitones) in ascending and descending forms while the other four end with one of two large intervals (9 or 10 semitones) in ascending and descending forms. Continuation tones consisted of the 15 diatonic tones in a two octave range centred on the final tone of the melodic context. The subjects were asked to rate how well the continuation tone continued the melody on a scale from 1 (extremely bad continuation) to 7 (extremely good continuation). The experiment yielded 120 continuation tone ratings for each subject. Significant inter-subject correlation for all subjects warranted the averaging of the data across subjects and training levels. The mean continuation tone ratings are available in Schellenberg (1996, Appendix

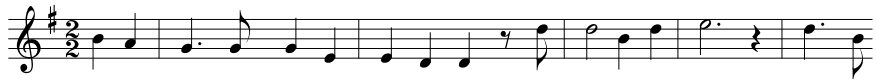
Fragment 1



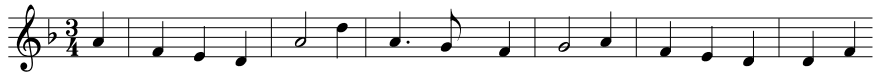
Fragment 2



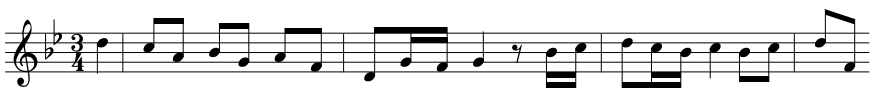
Fragment 3



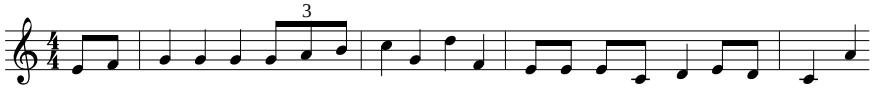
Fragment 4



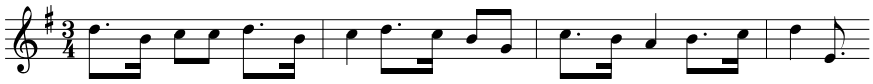
Fragment 5



Fragment 6



Fragment 7



Fragment 8

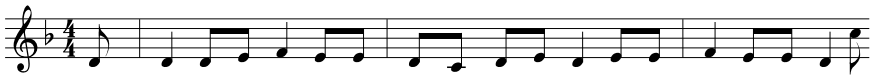


Figure 8.2: The melodic contexts used in Experiment 2 (after Schellenberg, 1996, Figure 3).

A).

The procedure was essentially the same as in Experiment 1 except that the statistical model returned distributions over an alphabet consisting of the diatonic tones an octave above and an octave below the final tone of each melodic fragment. Since the melodic fragments were longer, the short-term model was used in this experiment. Since it has been found that listeners are sensitive to short-term pitch distributional information in melodic material (Oram & Cuddy, 1995; Saffran *et al.*, 1999) and Schellenberg & Trehub (2003) have demonstrated accurate long-term pitch memory for familiar instrumental songs in ordinary listeners, it might be expected that regularities based on pitch height (e.g., `cpitch` or `cpitch-class`) will have an influence on expectation. Several

viewpoints, corresponding to hypotheses about the musical regularities underlying the observed patterns of expectation, were added to the set used in Experiment 1. In particular, it was hypothesised that melodic expectations might be influenced by tonality and the interaction of pitch with metric features. It is assumed that these score-based features are representative of perceived features and that the cognitive tasks of melodic segmentation (*e.g.,* Deliège, 1987; Ferrand *et al.*, 2002), tonality induction (Vos, 2000) and metre induction (*e.g.,* Eck, 2002; Toiviainen & Eerola, 2004) may be addressed independently from the present modelling concerns.

Regarding metric information, it was hypothesised that expectations might be influenced by regularities in pitch interval between notes occurring on metric pulses (`thrtactus`) and the interval of a note from the first note in the bar (`cpintfib`). In this case, the stimuli were presented to the subjects with a subtle pattern of emphasis in intensity based on the notated time signature (Schellenberg, 1996) in order to clarify the metrical structure (e.g., in the cases of Fragments 5 and 7 in Figure 8.2 which might otherwise be more naturally perceived in 2/4 metre). Regarding the effects of perceived tonality, it was hypothesised that expectations might be influenced by the representation of scale degree (`cpintfref`). The hypothesis underlying the use of statistical regularities in scale degree is closely related to an argument made by Krumhansl (1990) that the statistical usage of tones in existing musical traditions is the dominant influence on perceived tonal hierarchies (see §3.6). The viewpoint `cpintfref` was also linked with `dur`, `dur-ratio`, `ioi`, `cpint`, `cpintfip` and `fib` to investigate the interactions between perceived tonal structure and these dimensions of melodic, metric and rhythmic structure (see §5.4).

8.6.2 Results

The final multiple viewpoint system selected in this experiment enabled the statistical model to account for approximately 83% of the variance in the mean continuation tone ratings $[R = 0.910, R^2_{adj} = 0.827, F(1, 118) = 571.4, p < 0.001]$. The relationship between the patterns of expectation exhibited by the model and by the subjects in the experiments of Schellenberg (1996) is plotted with the fitted regression line in Figure 8.3. The statistical model provided a closer fit to the data than the two-factor model, which accounted for approximately 75% of the variance in the data $[R = 0.871, R^2_{adj} = 0.753, F(3, 116) = 121.9, p < 0.001]$, and this difference was found to be significant $[t(117) = 2.176, p < 0.05]$.

In order to examine the hypothesis that the statistical model subsumes the

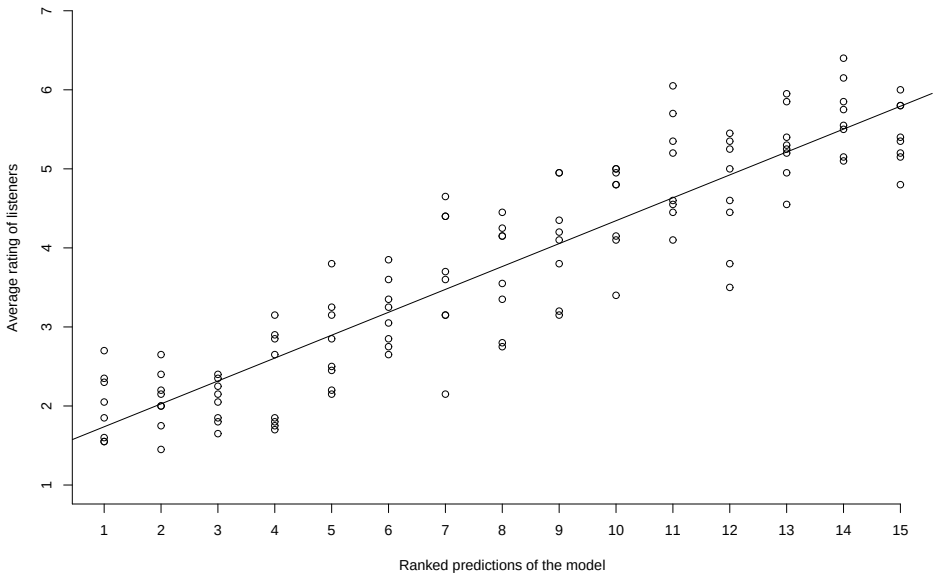


Figure 8.3: Correlation between subjects' mean goodness-of-fit ratings and the predictions of the statistical model for continuation tones in the experiments of Schellenberg (1996).

function of the bottom-up components of the two-factor model, a more detailed comparison of the two models was conducted. The expectations of the statistical model exhibit significant correlations in the expected directions with both components of the two-factor model: Proximity $[r(118) = -.738,\ p < 0.001]$; and Reversal $[r(118) = 0.489,\ p < 0.001]$. Furthermore, the fit of the statistical model to the behavioural data was not significantly improved by adding Proximity $[F(1, 117) = 3.865,\ p = 0.052]$ or Reversal $[F(1, 117) = 1.643,\ p = 0.203]$ to the regression model. However, adding both of these factors did significantly improve the fit of the statistical model to the data $[F(2, 116) = 6.034,\ p = 0.003]$. The resulting three-factor regression model accounted for approximately 84% of the variance in the mean continuation tone ratings $[R = 0.919,\ R^2_{adj} = 0.841,\ F(3, 116) = 210.7,\ p < 0.001]$.

Since the variables of the two-factor model are defined in terms of pitch interval, this departure from the results of Experiment 1 may reflect the relative paucity of features related to pitch interval selected in the present experiment (see Table 8.4). Since the feature selection algorithm does not cover the space of feature sets exhaustively, it is quite possible that there exist feature sets that include features related to pitch interval, that don't compromise the fit to the data achieved by the present statistical model but for which the addition of

Stage	Viewpoint Added	Viewpoint Dropped	R
1	`cpitch`		0.843
2	`cpintfib`		0.878
3	`cpintfip`		0.885
4	`cpintfref`⊗`cpint`		0.905
5	`cpitch`⊗`ioi`		0.909
6		`cpitch`	0.910

Table 8.4: The results of viewpoint selection in Experiment 2.

the two components of the two-factor model does not yield an improvement. Nonetheless, since the improvement yielded by the addition of the two predictors of the two-factor model was so small (an additional 1% of the variance, given 17% left unaccounted for by the statistical model alone), this analysis indicates that the statistical model *almost* entirely subsumes the function of Proximity and Reversal in accounting for the data collected by Schellenberg (1996).

Finally, in order to examine the selectivity of the two models, 50 sets of ratings for the stimuli ($N = 120$ for each set) were generated through random sampling from a normal distribution with a mean and SD equivalent to those of the listeners' ratings. With an alpha level of 0.05, just two of the 50 random vectors were fitted at a statistically significant level by each of the models and there was no significant difference between the fit of the two models for any of the 50 trials. Neither model is broad enough in its scope to successfully account for random data.

The results of viewpoint selection are shown in Table 8.4. Strong support was found for `cpitch` especially when linked with `ioi`, again illustrating the influence of joint regularities in pitch structure and rhythmic structure on expectations. The fact that `cpitch` was dropped immediately after the addition of `cpitch`⊗`ioi` suggests not only that the addition of the latter rendered the presence of the former redundant but also that regularities in `cpitch`, in the absence of rhythmic considerations, provide an inadequate account of the influence of pitch structure on expectations. In contrast to the impoverished contexts used in Experiment 1, the longer contexts used in this experiment are capable of invoking states of expectancy based on regularities in chromatic pitch structure. These regularities are likely to consist primarily of low-order intra-opus regularities captured by the short-term model although potentially higher-order extra-opus effects (via the long-term model) may also contribute since two of the training corpora contain Western folk melodies (*cf.* Krumhansl *et al.*, 1999).

The viewpoints `cpintfib` and `cpintfip` also contributed to improving the fit of the model to the human data, suggesting that regularities defined in reference to salient events (the first in the piece and the first in the current bar) are capable of exerting strong influences on melodic expectations.[6] Finally, one viewpoint representing a joint influence of regularities in tonal and melodic structure (`cpintfref`$\otimes$`cpint`) was selected. While this viewpoint improved the fit of the model, it is surprising that viewpoints modelling tonality were not selected earlier. This may be a result of the fact that British folk melodies are frequently modal (rather than tonal) and the fragments used do not always contain enough information to unambiguously specify the mode (A. Craft, personal communication, 9/9/2003).

Regularities in the four selected dimensions of existing melodies are such that the statistical model is able to exploit the longer contexts used in this experiment to provide a better account of the patterns of expectation observed in the experiment of Schellenberg (1996) than the two-factor model. A further illustration of the behaviour of the statistical model is presented graphically in Figures 8.4 and 8.5 which plot the responses of the model against those of the subjects in the experiment of Schellenberg (1996) to subsets of the continuation tone ratings. In both figures, the legend associates each of the eight melodic contexts with a distinct colour and also indicates the size (in semitones) and direction (ascending or descending) of the final implicative interval of the context. The plotted points are labelled with the size (in semitones) and direction (ascending or descending) of the realised interval formed by the continuation tone and are coloured according to the melodic context in which they appear.

Figure 8.4 plots the responses to continuation tones forming a small realised interval (five semitones or less according to the IR theory) with the final tone of the context. In accordance with the bottom-up principle of proximity, both the human listeners and the statistical model tend to exhibit high expectations for this set of small realised intervals and, within this set, larger intervals tend to receive lower ratings. However, both the model and the listeners exhibit plenty of exceptions to this trend. In the context of Fragment 5, for example, both the listeners and the model exhibit a relatively low expectation for a descending major second but a relatively high expectation for an ascending perfect fourth.

Figure 8.5 plots the responses to continuation tones for Fragment 4 which ends with a small ascending implicative interval (three semitones) and Fragment 6 which ends with a large ascending implicative interval (nine semitones). In accordance with the bottom-up principle of registral direction, both the hu-

[6] Note that the first note or the first note of the final bar of some, but not all, of the fragments is the tonic.

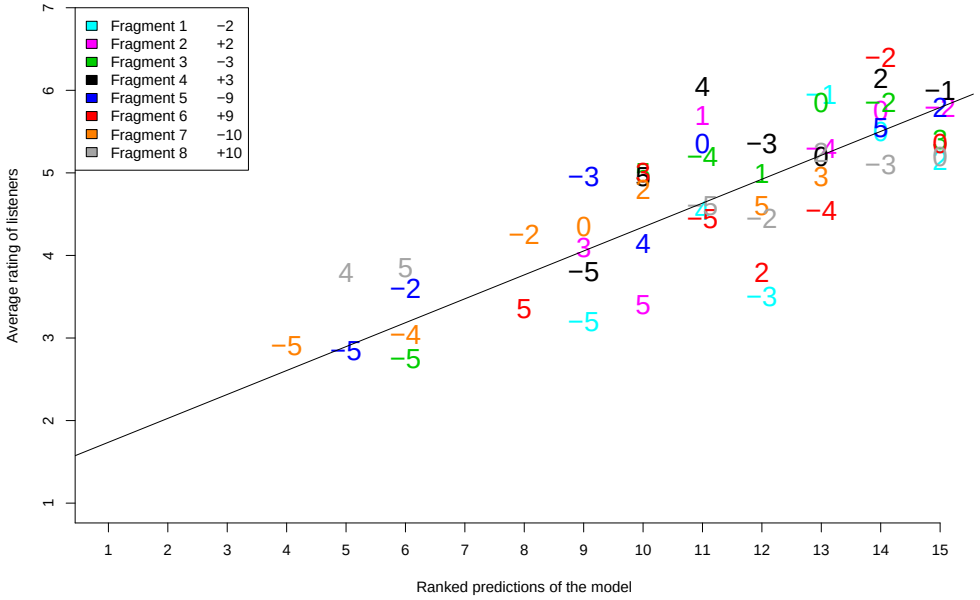


Figure 8.4: The relationship between the expectations of the statistical model and the principle of proximity (see text for details).

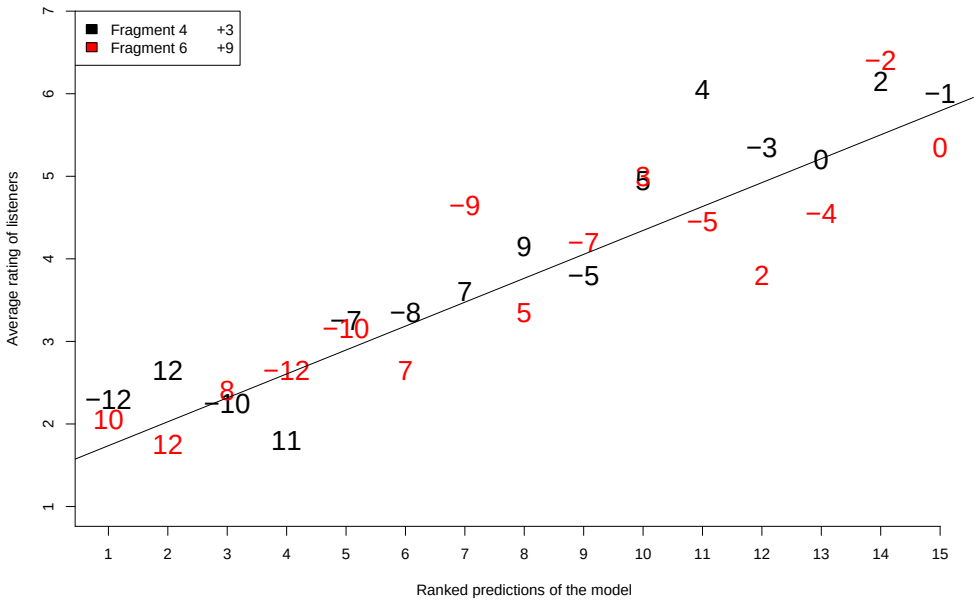


Figure 8.5: The relationship between the expectations of the statistical model and the principle of reversal (see text for details).

man listeners and the statistical model tend to expect the realised interval to maintain the registral direction of the (small) final implicative interval of Fragment 4 but to change the registral direction of the (large) final implicative interval of Fragment 6. As for the principle of proximity, both the listeners and the model exhibit exceptions to this rule such as an expectation for a return to the first tone of the implicative interval (which happens to be the tonic) in the case of Fragment 4 and an ascending step to the octave of the first tone of the implicative interval (which also happens to be the tonic) in the case of Fragment 6.

These examinations of the behaviour of the statistical model and the listeners demonstrate that the expectations of both tend to comply with the predictions of the two-factor model (although with a significant number of deviations). However, the fact that the statistical model yielded a closer fit to the behavioural data suggests that it provides, in addition, a more complete account of the manner in which the expectations of listeners deviate from the IR principles of proximity and pitch reversal. These observations indicate that the cognitive mechanisms responsible for the generation of melodic expectations can be accounted for largely in terms of the induction of statistical regularities in the musical environment which are approximately described by the principles of the two-factor model.

8.7 Experiment 3

8.7.1 Method

Most experimental studies of expectancy, including those of Cuddy & Lunny (1995) and Schellenberg (1996), have examined the responses of subjects only at specific points in melodic passages. Results obtained by this method, however, cannot address the question of how expectations change as a melody progresses (Aarden, 2003; Eerola *et al.*, 2002; Schubert, 2001; Toiviainen & Krumhansl, 2003). The purpose of this experiment was to examine the statistical model and the two-factor model (Schellenberg, 1997) in the context of expectations elicited throughout a melodic passage.

Manzara *et al.* (1992) have used an interesting methodological approach to elicit the expectations of listeners throughout a melody. The goal of the research was to derive an estimate of the entropy of individual pieces within a style according to the predictive models used by human listeners (see §6.2.2). The experimental methodology followed a betting paradigm developed by Cover & King (1978) for estimating the entropy of printed English. Subjects interacted

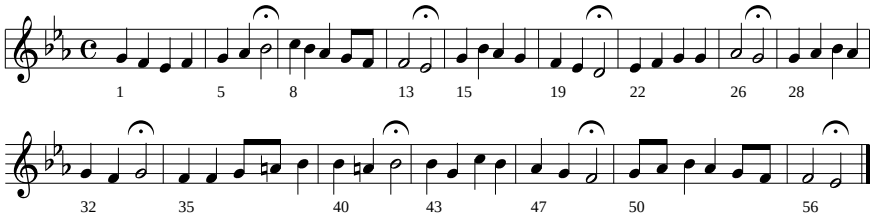
with a computer program presenting a score which retained all the information of the original except that the pitch of every note was B_4. Given an initial capital of $S_0 = 1.0$, the subjects were asked to move through the score sequentially, selecting the expected pitch of each note and betting a proportion p of their capital repeatedly until the selected pitch was correct, after which they could move to the next note. No time limits were set and the subjects could listen to the piece up to and including the current candidate note at any point. At each stage n, the subjects capital was incremented by $20pS_{n-1}$ if the selection was correct and decremented by the proportion bet if it was incorrect.[7] This proportional betting scheme was designed to elicit intuitive probability estimates for the next symbol to be guessed and rewards not only the correct guess but also accurate estimates of the symbol's probability. The entropy or *uncertainty* of a listener at stage n can be estimated as $\log_2 20 - \log_2 S_n$ where S_n is the capital won by the listener at this stage. Higher entropy indicates greater predictive uncertainty such that the actual pitch of the event is less expected.

Unlike the conventional probe tone method, the betting paradigm allows the collection of responses throughout a melodic passage (but see Toiviainen & Krumhansl, 2003, for a development of the probe tone methodology to allow the collection of real-time continuous responses). In addition, Eerola *et al.* (2002) report convergent empirical support for the use of entropy as a measure of predictability in melody perception (see §8.2.3.2). Furthermore, since it elicits responses prior to revealing the identity of the note and encourages the generation of probability estimates, the betting paradigm offers a more direct measure of expectation than the probe tone method. However, the responses of listeners in the betting paradigm are more likely to reflect the result of conscious reflection than in the probe tone paradigm and may be influenced by a potential learning effect.

The experimental stimuli used by Manzara *et al.* (1992) consisted of the melodies from Chorales 61 and 151 harmonised by J. S. Bach (Riemenschneider, 1941) which are shown in Figure 8.6. The subjects were grouped into three categories according to formal musical experience: novice, intermediate and expert. The experiment was organised as a competition in two rounds. Five subjects in each category took part in the first round with Chorale 151, while the two best performing subjects from each category were selected for the second round with Chorale 61. As an incentive to perform well, the overall winner in each of the categories won a monetary prize. The capital data for each event were averaged across subjects and presented as *entropy profiles* for

[7] There were 20 chromatic pitches to choose from.

61: *Jesu Leiden, Pein und Tod* (BWV 159)



151: *Meinen Jesum laß' ich nicht, Jesus* (BWV 379)

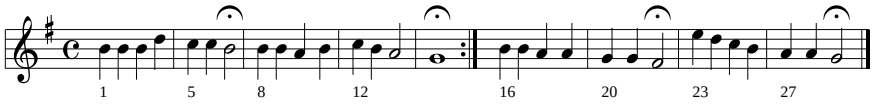


Figure 8.6: The two chorale melodies used in Experiment 3 (after Manzara *et al.*, 1992).

each chorale melody (see Figures 8.7 and 8.8).

Manzara *et al.* (1992) were able to make some interesting observations about the entropy profiles derived. In particular, it was found that the ultimate notes in phrases tended to be associated with lower uncertainty than those at the middle and beginning of phrases. High degrees of uncertainty, on the other hand, were associated with stylistically unusual cadential forms and intervals. The entropy profiles for both pieces also demonstrated high uncertainty at the beginning of the piece due to lack of context, followed by decreasing uncertainty as the growing context supported more confident predictions. For both pieces, the results demonstrated a rise in uncertainty near the end of the piece before a steep decline to the final cadence. Witten *et al.* (1994) found a striking similarity between the human entropy profiles and those generated by a multiple viewpoint statistical model derived from 95 chorale melodies (Conklin & Witten, 1995) suggesting that the relative degrees of uncertainty elicited by events throughout the pieces was similar for both the subjects and the model.

The experimental procedure used by Manzara *et al.* (1992) differs from that used by Cuddy & Lunny (1995) and Schellenberg (1996) as does the nature of the data collected. Consequently the methodology followed in this experiment differs slightly from those in Experiments 1 and 2. The main difference is that the expectations of the statistical model for each note in each melody were represented using entropy (the negative log of the estimated probability of the observed pitch). The performance metric was the regression coefficient of the mean entropy estimates for the subjects in the experiments of Manzara *et al.* (1992) regressed on the model entropy. Chorales 61 and 151 were not present

in the corpus of chorale melodies used to train the models; specifically, Chorale 151 was removed from Dataset 2 (see Table 4.1). Five viewpoints were added to the set used in Experiment 2 in order to examine the influence of phrase, metric and tonal structure on expectations elicited in the longer contexts of the two melodies. Specifically, viewpoints were incorporated that represent pitch interval between the first event in each consecutive bar (`thrbar`) and between events beginning and ending consecutive phrases (`thrfiph` and `thrliph`). A feature representing pitch in relation to the first note in the current phrase (`cpintfiph`) was also added to assess the potential influence of phrase level salience on expectations. Finally, a feature was added to represent whether a note is a member of the scale based on the notated key of the piece (`inscale`).

8.7.2 Results

The final multiple viewpoint system selected in this experiment enabled the statistical model to account for approximately 63% of the variance in the mean uncertainty estimates reported by Manzara *et al.* $[R = 0.796,\ R^2_{adj} = 0.629,\ F(1, 84) = 145,\ p < 0.001]$. Profiles for both model entropy and human entropy are shown in Figures 8.7 and 8.8 for Chorales 61 and 151 respectively. The entropy profiles illustrate the close correspondence between model uncertainty and human uncertainty throughout each of the chorale melodies (see also Witten *et al.*, 1994). The statistical model provided a closer fit to the data than the two-factor model, which accounted for approximately 13% of the variance in the data $[R = 0.407, R^2_{adj} = 0.134, F(3, 78) = 5.172, p < 0.01]$, and this difference was found to be significant $[t(79) = 5.15, p < 0.001]$. In the multiple regression analysis of the two-factor model and in comparing it to the statistical model, the data for the first two notes of each melody were not used since the two-factor model requires a context of a single interval in order to generate expectations.

In order to examine the hypothesis that the statistical model subsumes the function of the bottom-up components of the two-factor model, a more detailed comparison of the two models was conducted. The expectations of the statistical model exhibit a significant correlation in the expected direction with the Proximity component of the two-factor model $[r(80) = -.407,\ p < 0.001]$ but not with Reversal $[r(80) = 0.097,\ p = 0.386]$. Furthermore, the fit of the statistical model to the behavioural data was not significantly improved by adding Proximity $[F(1, 79) = 0.0122,\ p = 0.912]$, Reversal $[F(1, 79) = 0.0691,\ p = 0.793]$ or both of these factors $[F(2, 78) = 0.0476,\ p = 0.954]$ to the regression model. On this evidence, the statistical model entirely subsumes the

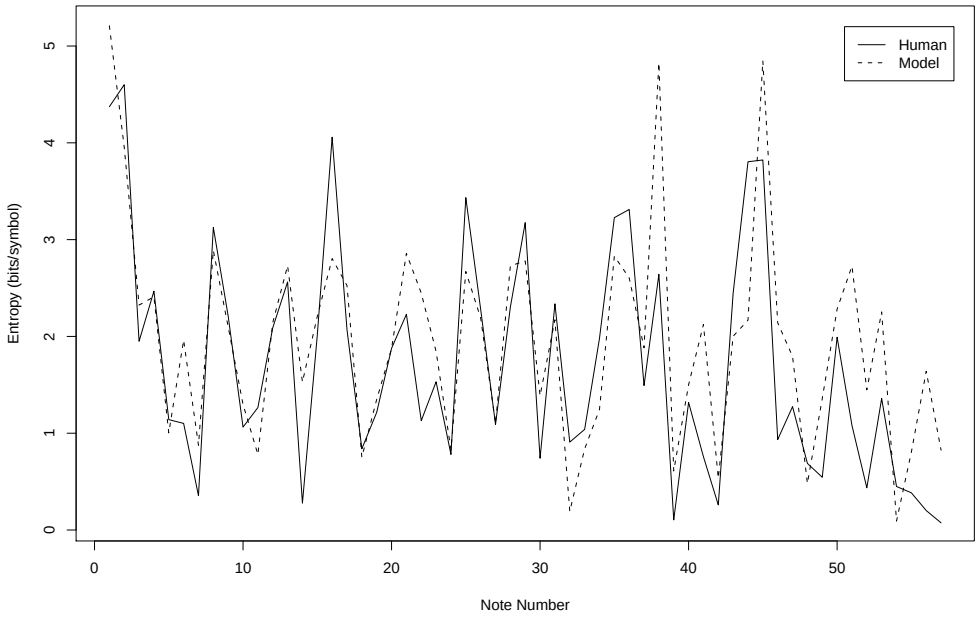


Figure 8.7: The entropy profiles for Chorale 61 averaged over subjects in the experiment of Manzara *et al.* (1992) and for the model developed in Experiment 3.

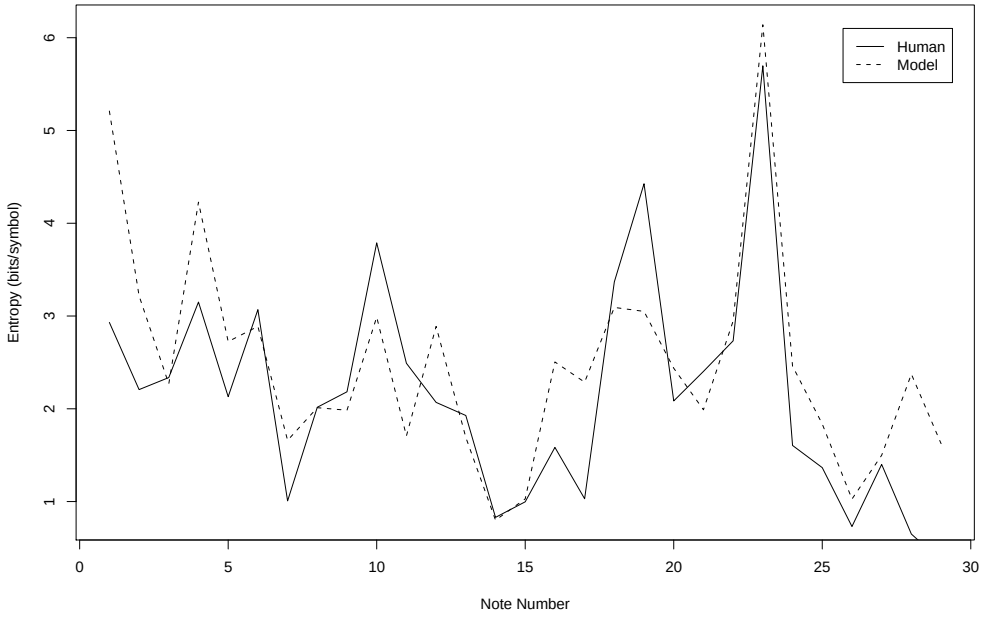


Figure 8.8: The entropy profiles for Chorale 151 averaged over subjects in the experiment of Manzara *et al.* (1992) and for the model developed in Experiment 3.

Stage	Viewpoint Added	R	H
1	`cpintfip`	0.737	2.293
2	`cpintfref`$\otimes$`dur-ratio`	0.794	2.162
3	`thrfiph`	0.796	2.143

Table 8.5: The results of viewpoint selection in Experiment 3.

function of Proximity and Reversal in accounting for the data collected by Manzara *et al.* (1992).

Finally, in order to examine the selectivity of the two models, 50 sets of entropy estimates for the two chorales were generated through random sampling from a normal distribution with a mean and SD equivalent to those of the listeners' entropy estimates. With an alpha level of 0.05, just two of the 50 random vectors were fitted at a statistically significant level by the two factor model and in only one of these trials was there a significant difference between the fit of the two models. Neither model is broad enough in its scope to successfully account for random data.

The results of viewpoint selection are shown in Table 8.5. As in Experiments 1 and 2, `cpintfip` made a strong contribution to the fit of the model. Support was also found for one linked viewpoint representing the influence of tonality (`cpintfref` $\otimes$ `dur-ratio`) and fact that this viewpoint was selected over its primitive counterpart again provides evidence for the interactive influence of rhythmic and pitch structure on expectancy. Finally, some support was found for an influence of phrase level regularities on expectancy (`thrfiph`).

In addition to showing the regression coefficient (R) which was used as the evaluation metric in the viewpoint selection experiment, Table 8.5 also shows the entropy of the model averaged over all events considered during prediction of the two melodies (H). The observation that H decreases as R increases suggests a rational cognitive basis for the selection of melodic features in the generation of expectations: features may be selected to increase the perceived likelihood (or expectedness) of events and reduce redundancy of encoding (Chater, 1996, 1999). In order to examine this hypothesis, a further selection experiment was run in which viewpoints were selected to minimise model uncertainty (as measured by mean per-event entropy) over Chorales 61 and 151. The results of this experiment are shown in Table 8.6 which shows average model uncertainty (H) and the regression coefficient (R) of the mean entropy estimates of the subjects in the experiments of Manzara *et al.* (1992) regressed on the model entropy for each selected system.

Once again, the viewpoint selection results generally exhibit an inverse

Stage	Viewpoint Added	R	H
1	cpintfref⊗cpint	0.657	2.061
2	cpint⊗dur	0.693	1.972
3	cpintfip	0.738	1.937
4	cpintfref⊗fib	0.750	1.916
5	thrfiph	0.759	1.903

Table 8.6: The results of viewpoint selection for reduced entropy over Chorales 61 and 151 in Experiment 3.

trend between R and H. However, while the systems depicted in Tables 8.5 and 8.6 show a degree of overlap, Table 8.6 also reveals that exploiting regularities in certain features (especially those related to melodic interval structure) improves prediction performance but does not yield as close a fit to the behavioural data as the system shown in Table 8.5. A closer inspection of all 247 multiple viewpoint systems considered in this experiment revealed a significant negative correlation between R and H for values of H greater than 2.3 bits/symbol $[r_s(N = 45) = -0.85, p < 0.001]$ but not below this point $[r_s(N = 202) = -0.05, p = 0.46]$. If listeners do focus on regularities in melodic features so as to reduce uncertainty, this relationship may be subject to other constraints such as the number and kind of representational dimensions to which they can attend concurrently.

8.8 Discussion and Conclusions

The first goal defined in this chapter was to examine whether models of melodic expectancy based on statistical learning are capable of accounting for the patterns of expectation observed in empirical behavioural research. The sequence learning system developed in Chapters 6 and 7 and the two-factor model of expectancy (Schellenberg, 1997) were compared on the basis of scope, selectivity and simplicity (Cutting *et al.*, 1992; Schellenberg *et al.*, 2002). The two models could not be distinguished on the basis of selectivity since neither was found to account for random patterns of expectation in any of the three experiments. Regarding the scope of the two models, the results demonstrate that the statistical model accounted for the behavioural data as well as, or better than, the two-factor model in all three of the reported experiments. Furthermore, the difference between the two models became increasingly apparent when expectations were elicited in the context of longer and more realistic melodic contexts (see also Eerola *et al.*, 2002). Finally, regarding the simplicity of the

two models, the results indicate that the statistical model entirely (or almost entirely in the case of Experiment 2) subsumes the function of the principles of Proximity and Reversal (Schellenberg, 1997) in accounting for the expectations of listeners, rendering the inclusion of these rules in an additional system of innate bottom-up predispositions unnecessary.

Altogether, these experimental results demonstrate that patterns of expectation elicited in a range of melodic contexts can be accounted for in terms of the combined influence of sensitivities to certain dimensions of the musical surface, relatively simple learning mechanisms and the structure of the musical environment. In contrast to one of the central tenets of the IR theory, universal symbolic rules need not be assumed to account for experimentally observed patterns of melodic expectation. The quantitatively formulated bottom-up and top-down principles of the IR models may be viewed as formalised approximations to behaviour that emerges as a result of statistical induction of regularities in the musical environment achieved by a single cognitive system (*cf.* Thompson & Stainton, 1998).

The second goal in this chapter was to undertake a preliminary examination of the kinds of melodic feature that afford regularities capable of supporting the acquisition of the observed patterns of expectation. In each experiment, only a small number of features (three in Experiments 1 and 3, and four in Experiment 2) were selected by the forward stepwise selection procedure even though the evaluation functions used did not explicitly penalise the number of features used by the statistical model. In all three experiments, it was found that regularities in pitch structure defined in relation to the first note in a melody are capable of exerting strong influences on expectancy. This influence of primacy on perceived salience suggests that the first note in a melody provides a strong reference point with which subsequent structures are compared in the generation of expectations (Cohen, 2000; Cuddy & Lunny, 1995; Longuet-Higgins & Steedman, 1971; Thompson *et al.*, 1997). Furthermore, the results of all three experiments provide evidence that expectations are influenced by regularities in the interaction of pitch structure and rhythmic structure (see also Jones, 1987; Jones & Boltz, 1989).

In addition, the experimental results suggest that induced regularities in different melodic features may influence expectancy to varying degrees in different contexts. The short contexts in Experiment 1, for example, tended to generate expectations based on regularities in melodic interval structure rather than chromatic pitch structure. In the second experiment, on the other hand, support was found for the influence of chromatic pitch structure as well as met-

ric structure and tonal regularities. Finally, in Experiment 3, support was found for the influence of tonal structure and phrase-level salience on the generation of expectations. These differences suggest that melodic contexts differ in the extent to which they emphasise different features used in cuing attention to salient events. The results of Experiment 3 also provide some evidence for a relationship, across different feature sets, between the predictive uncertainty of the statistical model and its fit to the behavioural data suggesting that, subject to other constraints, listeners employ representations which increase the perceived likelihood of melodic stimuli (Chater, 1996, 1999). The mechanisms by which attention is drawn to different features in different melodic contexts and how regularities in these dimensions influence expectancy is an important topic for future empirical research. Improved methodologies for eliciting and analysing continuous responses to music (Aarden, 2003; Eerola *et al.*, 2002; Schubert, 2001; Toiviainen & Krumhansl, 2003) will form an important element in this research.

The experimental results provide support for the present theory of melodic expectation in terms of the influence of melodic context on the invocation of learnt regularities. In particular, the results confirm that regularities in existing melodic traditions are sufficient to support the acquisition of observed patterns of expectation. According to the theory, expectations will also be subject to the influence of prior musical experience. Future research should examine this aspect of the theory in greater depth. It would be predicted, for example, that a model exposed to the music of one culture would predict the expectations of people of that culture better than a model trained on the music of another culture and *vice versa* (see also Castellano *et al.*, 1984). The theory also predicts that observed patterns of expectation will become increasingly systematic and complex with increasing age and musical exposure (*cf.* Schellenberg *et al.*, 2002). Future research might examine the developmental profile of expectations exhibited by the model as it learns, yielding testable predictions about developmental trajectories in the acquisition of melodic expectations exhibited by infants (see also Plunkett & Marchman, 1996).

Another fruitful avenue for future research involves a more detailed examination of the untested assumptions of the model, the elaboration of the theory and the proposition of hypotheses at finer levels of detail (Desain *et al.*, 1998). Such hypotheses might concern, for example, the developmental status of the features assumed to be present in the musical surface and the derivation of other features from this surface as well as how the interaction between the long- and short-term models is related to the effects of intra- and extra-opus

experience. The examination of expectations for more complex musical structures embedded in polyphonic contexts may reveal inadequacies of the model. For example, its reliance on local context in generating predictions may prove insufficient to account for the perception of non-local dependencies and recursively embedded structure (Lerdahl & Jackendoff, 1983). Conversely, the computational model may be overspecified in some regards as a model of human cognition. For example, schematic influences on expectancy are likely to be subject to the effects of limitations on working memory although the model is not constrained in this regard (Reis, 1999).

To conclude, not only does the theory put forward in this chapter provide a compelling account of existing data on melodic expectancy: it also makes a number of predictions for future research. In this regard, the modelling strategy followed in the present research constitutes a rich source of new hypotheses regarding the influence of musical context and experience on expectations and provides a useful framework for the empirical examination of these hypotheses.

8.9 Summary

In this chapter, the predictive system presented in Chapters 6 and 7 was applied to the modelling of expectancy in melody perception. In §8.2, the concept of expectancy in music was introduced and the influential theoretical accounts of Meyer (1956, 1967, 1973) and Narmour (1990, 1992) were reviewed. Empirical support for these theoretical accounts suggests that while the IR theory can account well for observed patterns of expectation, research to date does not support the rigid distinction between bottom-up and top-down components of the theory nor the hypothesised innateness of the bottom-up principles. In §8.3, an alternative theory of melodic expectancy was presented according to which the observed patterns of expectation can be accounted for in terms of the induction of statistical regularities in existing corpora of music. Patterns of expectation which do not vary between musical styles are accounted for in terms of simple regularities in music which may be related to the constraints of physical performance. The computational system developed in Chapters 6 and 7 was discussed in terms of this theory and the experimental methodology used to examine the behaviour of the system as a model of melodic expectancy was introduced in §8.4. The experimental design and results of three experiments with increasingly complex melodic stimuli were presented and discussed in §8.5, §8.6 and §8.7 respectively. Finally, a general discussion of the experimental results and their implications was presented in §8.8.

CHAPTER 9

MODELLING MELODIC COMPOSITION

9.1 Overview

The goal in this chapter is to examine, at the computational level, the intrinsic demands of the task of composing a successful melody. Of particular concern are constraints placed on the representational primitives and the expressive power of the compositional system. In order to achieve these goals, three of the multiple viewpoint systems developed in previous chapters are used to generate novel pitch structures for seven of the chorale melodies in Dataset 2. In §9.3, null hypotheses are presented which state that each of the three models is consistently capable of generating chorale melodies which are rated as equally successful, original and creative examples of the style as the original chorale melodies in Dataset 2. In order to examine these hypotheses, experienced judges are asked to rate the generated melodies together with the original chorale melodies on each of these three dimensions. The experimental methodology, described in §9.4, is based on the Consensual Assessment Technique developed to investigate psychological components of human creativity. In §9.2.3 and §9.2.4 it is argued that this methodology addresses some notable limitations of previous approaches to the evaluation of computational models of compositional ability. The results, presented in §9.5, warrant the rejection of the null hypothesis for all three of the systems. In spite of steps taken to address the limitations of previous context modelling approaches to generating music, the finite context grammars making up these systems show little ability to meet the computational demands of the task regardless of the representa-

tional primitives used. Nonetheless, a further analysis identifies some objective features of the chorale melodies which exhibit significant relationships with the ratings of stylistic success. These results, in turn, suggest ways in which the computational models fail to meet intrinsic stylistic constraints of the chorale genre. Adding viewpoints to the multiple viewpoint systems to address these concerns significantly improves the prediction performance of these systems.

9.2 Background

9.2.1 Cognitive Modelling of Composition

In striking contrast to the amount of cognitive-scientific research which has been carried out on music perception, cognitive processes in composition remain largely unexamined (Baroni, 1999; Sloboda, 1985). This section contains a review of research which has been carried out on the cognitive modelling of music composition with an emphasis on computational approaches. It should be noted that this review intentionally excludes research on the computer generation of music in which cognitive-scientific concerns in the construction and evaluation of the computational models are not apparent.

Given the current state of knowledge about cognitive processes in composition, Johnson-Laird (1991) in his study of jazz improvisation (see §3.3) argues that it is fundamental to understand what the mind has to compute in order to generate an acceptable improvisation before examining the precise nature of the algorithms by which it does so (see §2.4).[1] In order to study the intrinsic constraints of the task, Johnson-Laird applied grammars of varying degrees of expressive power to different subcomponents of the problem. The results of this analysis suggest that while a finite state grammar is capable of computing the melodic contour, onset and duration of the next note in a jazz improvisation, its pitch must be determined by constraints derived from a model of harmonic movement which requires the expressive power of a context free grammar.

Lerdahl (1988a) explores the relationship between perception and composition and outlines some cognitive constraints that this relationship places on the cognitive processes of composition. Lerdahl frames his arguments within a context in which a *compositional grammar* generates both a structural description of a composition and, together with intuitive perceptual constraints, its realisation as a concrete sequence of discrete events which is consumed by

[1] Improvisation may be regarded as a special case of composition in which the composer is also the performer and is subject to additional constraints of immediacy and fluency (Sloboda, 1985).

a *listening grammar* that, in turn, yields a perceived structural description of the composition. A further distinction is made between *natural* and *artificial* compositional grammars: the former arise spontaneously within a culture and are based on the listening grammar; the latter are consciously developed by individuals or groups and may be influenced by any number of concerns. Noting that both kinds of grammar coexist fruitfully in most complex and mature musical cultures, Lerdahl argues that when the artificial influences of a compositional grammar carry it too far from the listening grammar, the intended structural organisation can bear little relation to the perceived structural organisation of a composition. Lerdahl (1988a) goes on to outline a number of constraints placed on compositional grammars via this need for the intended structural organisation to be recoverable from the musical surface by the listening grammar. These constraints are largely based on the preference rules and stability conditions of GTTM (see §3.3).

The proposal that composition is constrained by a mutual understanding between composers and listeners of the relationships between structural descriptions and the musical surface is expanded by Temperley (2003) into a theory of *communicative pressure* on the development of musical styles. Various phenomena are discussed within the context of this theory including the relationship between the traditional rules of voice leading and principles of auditory perception (Huron, 2001) as well as trade-off relationships between syncopation and rubato in a range of musical styles.

Baroni (1999) also discusses grammars for modelling the cognitive processes involved in music perception and composition basing his arguments on his own development, implementation and use of grammars for the structural analysis of a number of musical repertoires (Baroni *et al.*, 1992). Baroni characterises a listening grammar as a collection of morphological categories which define sets of discrete musical structures at varying levels of description and a collection of syntactical rules for combining morphological units. He argues that such a grammar is based on a stylistic mental prototype acquired through extensive exposure to a given musical style. While the listening grammar is largely implicit, according to Baroni, the complex nature of composition requires the acquisition of explicit grammatical knowledge through systematic, analytic study of the repertoire. However, Baroni (1999) states that the compositional and listening grammars share the same fundamental morphology and syntax. The distinguishing characteristics of the two cognitive activities lie in the technical procedures underlying the effective application of the syntactical rules. As an example, Baroni examines hierarchical structure in the listening

and compositional grammars: for the former, the problem lies in picking up cues for the application of grammatical rules and anticipating their subsequent confirmation or violation in a sequential manner; for the latter, the structural description of a composition may be generated in a top-down manner.

9.2.2 Music Generation from Statistical Models

Conklin (2003) discusses the generation of music using statistical models from the perspective of generating pieces which have high probability of occurring according to the model. Conklin examines four methods for generating compositions from statistical models of music. The first and simplest is sequential random sampling where an event is sampled from the estimated distribution of events at each sequential event position. The sampled event is appended to the generated piece and the next event is sampled until a specified limit on the length of the piece is reached. Since events are generated in a random walk, there is a danger of straying into local minima in the state space of possible compositions. More importantly, however, this method suffers from the fact that it greedily tends to generate high probability events without regard for the overall probability of the generated piece. Events with high estimated probabilities generated at one stage may constrain the system at a later stage to generate a piece with a low overall probability. Nonetheless, most efforts to generate music from statistical models, including all those discussed in §3.4, have used this method.

One statistical modelling technique which addresses these problems is the Hidden Markov Model (HMM) which generates observed events from hidden states (Rabiner, 1989). Training a HMM involves adjusting the probabilities conditioning the initial hidden state, the transitions between hidden states and the emission of observed events from hidden states so as to maximise the probability of a training set of observed sequences. A trained HMM can be used to estimate the probability of an observed sequence of events and to find the most probable sequence of hidden states given an observed sequence of events. The latter task can be achieved efficiently for a first-order HMM using a dynamic programming solution known as the Viterbi algorithm and a similar algorithm exists for first-order (visible) Markov models. Allan (2002) has used the Viterbi algorithm to generate the most likely sequence of underlying harmonic states given an observed chorale melody. Furthermore, Allan demonstrates that this method is capable of generating significantly more probable harmonic progressions than are typically obtained using sequential random sampling.

In the context of complex statistical models, such as those developed in

Chapters 6 and 7, the Viterbi algorithm suffers from two problems (Conklin, 2003). First, its time complexity increases exponentially with the context length of the underlying Markov model. Second, it is difficult to formulate such models in an appropriate manner for using the Viterbi algorithm.

There do exist tractable methods for sampling from complex statistical models which address the limitations of random sampling (Conklin, 2003). The *Metropolis-Hastings algorithm* is a Markov Chain Monte Carlo (MCMC) sampling method which provides a good example of such techniques (MacKay, 1998). The following description applies the Metropolis-Hastings algorithm within the framework developed in Chapters 6 and 7. Given a trained multiple viewpoint system m for some basic type τ_b, in order to sample from the target distribution $p_m(s \in [\tau_b]^*)$, the algorithm constructs a Markov chain in the state space of possible viewpoint sequences $[\tau_b]^*$ as follows:

1. set the iteration number $k = 0$, the desired number of iterations $N =$ some large value; the initial state $s_0 =$ some viewpoint sequence $t_1^j \in [\tau_b]^*$ of length j;

2. select an event index $1 \leq i \leq j$ either at random or based on some ordering of the indices;

3. let s'_k be the sequence obtained by replacing the event t_i at index i of s_k with a new event t'_i sampled from a proposal distribution q which may depend on the current state s_k – in the present context, an obvious choice for q would be $\{p_m(t|t_1^{i-1})\}_{t \in [\tau_b]}$;

4. accept the proposed sequence with probability:

$$\min\left[1,\ \frac{p_m(s'_k\ )\ \cdot\ q(t_i)}{p_m(s_k)\ \cdot\ q(t'_i)}\right];$$

5. if accepted, set $s_{k+1} = s'_k$, else set $s_{k+1} = s_k$;

6. if $k < N$, set $k = k + 1$ and return to step 2, else return s_k.

If N is large enough, the resulting event sequence s_{N-1} is guaranteed to be an unbiased sample from the target distribution $p_m([\tau_b]^*)$. However, there is no general theoretical method for assessing the convergence of MCMCs nor to estimate the number of iterations required to obtain an unbiased sample (MacKay, 1998). Another popular MCMC method, *Gibbs sampling* corresponds to a special case of Metropolis sampling in which the proposal density q is the full distribution $\{p_m(s'_k)\}_{t \in [\tau_b]}$ and the proposal is always accepted. Since this

distribution may be expensive to compute, Gibbs sampling can add significant computational overheads to the sampling procedure. Finally, because these sampling algorithms explore the state space using a random walk, they still suffer from the problem of falling into local minima. The effects of this limitation may be counteracted to some extent by selecting a high probability event sequence as the start state (Conklin, 2003).

Another solution to the problems resulting from the use of random walks is to introduce symbolic constraints on the generation of events. Hall & Smith (1996), for example, placed structural constraints on harmonic movement and the metric positions of rhythmic groups to prevent the generation of stylistically uncharacteristic features in their statistical model of blues tunes. In a similar vein, Hild *et al.* (1992) employed symbolic voice leading constraints in their neural network model of chorale harmonisation. While these studies derived their constraints from stylistic analyses or music-theoretic concerns, Povel (2004) describes a system for generating melodies based on perceptual constraints. Research on the perception of rhythm, harmony and contour are used to constrain the production of tonal melodies in order to examine whether such constraints are necessary and sufficient determinants of tonal structure.

All the methods discussed so far generate new pieces through the substitution of single events. As a consequence, they are unlikely to provide a satisfactory model of phrase or motif level structure and, in particular, to preserve the structure of repeated phrases or variations. Although the short term model (see §6.2.4) is intended to provide a model of intra-opus structure, it still relies on single-event substitutions. In order to address these concerns, Conklin (2003) argues that pattern discovery algorithms (*e.g.,* Cambouropoulos, 1998; Conklin & Anagnostopoulou, 2001) may be used to reveal phrase level structure at various degrees of abstraction which may subsequently be preserved during stochastic sampling. The discovery and use of motif classes in generating melodic variations by Hörnel (1997) is an example of this approach (see §3.5).

9.2.3 Evaluating Computational Models of Composition

Analysis by synthesis is a method for evaluating computational models of music by generating compositions which are subsequently analysed with respect to the objectives of the implemented model. This method of evaluation has a long history and it has been argued that one of the primary advantages of a computational approach to the analysis of musical styles is the ability to generate new pieces from an implemented theory for the purposes of evaluation (Ames &

Domino, 1992; Camilleri, 1992; Sundberg & Lindblom, 1976, 1991). However, the evaluation of the generated music raises methodological issues which have typically compromised the benefits potentially afforded by the computational approach.

In many cases, the compositions are evaluated with a single subjective comment such as: "[the compositions] are realistic enough that an unknowing listener cannot discern their artificial origin" (Ames & Domino, 1992, p. 186); "[the program] seems to be capable of producing musical results" (Ebcioğlu, 1988, p. 49); or "The general reactions of Swedes listening to these melodies informally are that they are similar in style to those by Tegnér" (Sundberg & Lindblom, 1976, p. 111). Johnson-Laird (1991, p. 317) simply notes that "The program performs on the level of a competent beginner" and gives an informal account of how the program undergenerates and overgenerates. In short, "the gap between the cognitive work, that brings forward these models, and the experimental approach, that should validate them, is huge" (Desain *et al.*, 1998, p. 154). This lack of emphasis on evaluation has the effect of making it very difficult to compare and contrast different theories intersubjectively.

Other research has used expert stylistic analyses to evaluate the compositions generated by computational systems. This is possible when a computational model is developed to account for some reasonably well defined stylistic competence or according to critical criteria derived from music theory or research in music psychology. For example, Ponsford *et al.* (1999) gave an informal stylistic appraisal of the harmonic progressions generated by their n-gram models. A more intersubjective method of appraisal is described by Hild *et al.* (1992) who developed a system which would harmonise in the style of J. S. Bach. The harmonisations produced by their system were judged by music professionals to be on the level of an improvising organist. A more detailed appraisal of computer generated harmonies was obtained by Phon-Amnuaisuk *et al.* (1999) who had the generated harmonisations evaluated by a university senior lecturer in music according to the criteria used for examining first year undergraduate students' harmony.

However, even when stylistic analyses are undertaken by groups of experts, the results obtained are typically still qualitative in nature. For a fully intersubjective analysis by synthesis, the evaluation of the generated compositions should be quantitative. One possibility is to use an adapted version of the Turing test in which subjects are presented with pairs of compositions (of which one is computer generated and the other human composed) and asked to state which they believe to be the computer generated composition (Marsden, 2000).

This approach has been adopted by Triviño-Rodriguez & Morales-Bueno (2001) and Hall & Smith (1996) for the evaluation of, respectively, chorale melodies and blues melodies generated from statistical models of music (see §3.4).

Hall & Smith (1996) randomly selected ten blues melodies from their corpus, removed them from the training set and used the remaining 48 melodies to train their model. Twenty blues melodies were generated from the model and ten of these were randomly selected for evaluation. The original and generated melodies were randomly assembled into ten pairs each with a fixed, randomly generated presentation order. These pairs of stimuli were presented in fixed order to 198 subjects of whom only 23 had more than four years of training on a musical instrument. A Chi-squared test was performed for the numbers of listeners making less than two errors and two or more errors. The results suggested that the subjects were unable to reliably distinguish the human composed and computer generated melodies.

Triviño-Rodriguez & Morales-Bueno (2001) adopted a similar experimental procedure in which 52 listeners were presented with two pairs of computer generated and human composed chorale melodies. Triviño-Rodriguez & Morales-Bueno do not provide details of the selection of the chorale melodies in either category nor of the presentation orders used in the experiment. The results showed that the listeners were able to correctly classify melodies in 55% of trials. A more informal variant of this approach (called The Game) has also been used to evaluate computer generated compositions by Cope (2001, see §3.3), who reports that listeners typically average between 40% and 60% correct responses.

These musical Turing tests have the advantage of yielding empirical, quantitative results which may be appraised intersubjectively. Overall, they have demonstrated the inability of subjects to reliably distinguish between computer generated and human composed compositions. However, the methodology suffers from a number of difficulties all of which stem from its failure to examine the criteria being used to judge the compositions.

First, it must be noted that these studies have not typically used musically trained subjects capable of distinguishing important stylistic features of the compositions. The problem of the relationship between objective properties of artefacts and subjective judgements of their aesthetic value has a long history in philosophy, dating back at least as far as David Hume's classic essay on taste (Hume, 1965), first published in 1757. Hume argued that there exists some universality in the relation between the attributes of objects and aesthetic experience and that this permits intersubjective agreement on aesthetic issues.

This universal standard of taste, however, may often be distorted by lack of culturally embedded experience in discriminating the relevant properties of objects that may elicit an aesthetic response.

To consider a musical example, Stobart & Cross (2000) report an analysis of the music of the indigenous people of Northern Potosí in Bolivia which underlines the importance of culture and language in shaping musical preferences. Rather than viewing relatively long events as having metrical salience (as Western listeners typically do, see §5.4.1), the Potosíans take the first event of a piece as initiating the metrical framework. Stobart & Cross (2000) suggest that this may be related to the fact that the only fixed position prosodic stress in their language occurs on the first syllable of a word.[2] Clearly, the subjective appraisals of Western listeners on initial exposure to Potosían music would be subject to the distorting effects of a likely misinterpretation of the metric framework. Studies such as this emphasise the importance of cultural and stylistic experience in the appraisal of music and suggest that judges of computer generated music should possess considerable knowledge of the target style.

A second and less easily remediable problem with the Turing test approach is that the paradigm used is likely to have the effect of shifting attention towards searching for musical features expected to be generated by a computer model rather than concentrating on stylistic features of the compositions. Drawing once again on philosophical accounts of aesthetic judgement, Kant (1952) proposed that two individuals can be said to be perceiving the same object only when they possess the same faculties of perception and understanding, which operate identically in both cases – *i.e.,* their cognitive representation of the object is the same. Kant defined a purely aesthetic judgement (in contrast to moral and pragmatic judgements) in stringent terms as an assessment of the perceived form of an artefact which is free of any interest (desires, needs and so on) that the judge may have in the actual existence of the artefact.

It seems likely that the Turing test would stimulate an emotional interest in most judges in terms of, for example, competitive intellectual pride (see Cope, 2001, ch. 2). This vested interest may in turn distort the assessment by causing the judges to rely on preconceived notions of the capacities of computers rather than on their knowledge and appreciation of the musical style. For example, in an evaluation of human composed and computer generated rhythmic patterns, by Pearce & Wiggins (2001), in which subjects were asked to state whether

[2]Using a similar argument, Hall (1953) suggests that the reason that Elgar's popularity was restricted to England (at the time), while English composers such as Britten and Vaughan-Williams were well liked abroad, was due to the manner in which his music reflects the wide pitch range and predominantly descending patterns of intonation peculiar to British English (as compared with American English and most Continental languages).

they thought a given pattern was created by a human or a computer, there was a systematic bias towards classifying patterns as computer generated. The informally collected comments of the subjects suggested that they perceived the task as a challenge to catch out the computer which may have contributed to the bias.

A final and more serious shortcoming of the Turing test methodology is that it fails to shed light on the fundamental musicological and psychological questions of interest: Which stylistic features present or absent in the computer generated compositions facilitate discrimination performance? Which cognitive or stylistic hypotheses embodied in the generative system influence the judgements of listeners and how do they do so? Cope (2001, p. 21) suggests that listeners who perform well in The Game should try to "identify those characteristics which gave the machine-composed examples away" while listeners who performed poorly should "try to discover what led the machine-composed examples to sound as if they were human-composed." Ideally, we would like to formalise these suggestions for qualitative responses into a concretely specified, quantitative and empirical methodology for the detailed examination, reformulation and elaboration of computational models of musical competence (see §2.4 and §2.5).

9.2.4 Evaluating Human Composition

One approach to developing alternatives to the Turing test paradigm discussed in §9.2.3 is to examine empirical methodologies used to evaluate human compositional ability. The vast majority of existing methodologies of this kind have been developed to assess the compositional creativity of school children and students. As noted by Auh (2000), such methodological approaches differ on at least two different dimensions. The first dimension concerns the kind of data collected which may be qualitative or quantitative. Since the present concern is with quantitative methodologies, qualitative approaches such as those of Folkestad *et al.* (1990) are not considered here (see Sloboda, 1985, ch. 4, for a review). The second dimension identified by Auh (2000) concerns the object of the evaluation which may be the creative individuals themselves, the process of creation, the product or artefact resulting from this process or the environment within with the creative individual operates. This taxonomy of objects of study is quite common in the wider field of research on creativity (Brown, 1989; Jackson & Messick, 1965; Plucker & Renzulli, 1999).

In terms of the composition of music, the creative individual is the composer and the product is the composition. Attempts to measure characteristics

of composers have focused on specific musical factors (such as musical achievement and experience), general factors (such as academic achievement, IQ and gender) as well as cognitive style and personality traits (Auh, 2000). Studies of the creative environment examine such factors as the social, cultural, occupational, political and economic contexts within which a composer works (see *e.g.,* Simonton, 1998). Research on the process of composition has typically focused on the analysis of observed sessions performing some well defined compositional task into categories on the basis of time spent performing a given activity, the working score, recorded verbal utterances and the final composition (Colley *et al.*, 1992; Davidson & Welsh, 1988; Folkestad *et al.*, 1990; Kratus, 1989, 1994; Webster, 1987).

Of most relevance to the present research, however, are methodologies which have been developed to evaluate the final product of the compositional process. Theoretical approaches to evaluating the creativity of products typically stress the need to assess both the originality, novelty or unusualness and the appropriateness or value of the artefact (Amabile, 1996; Boden, 1990; Jackson & Messick, 1965; Mayer, 1999; Ritchie, 2001). The unusualness of a response can only be judged in relation to a set of *norms*, based on a group of existing artefacts, which serve as a judgemental standard. The judgemental standard for evaluating appropriateness is the *context* which reflects both external factors, which must be interpreted logically (*i.e.,* in terms of the demands of the task) and psychologically (*i.e.,* in terms of the intentions of the creator), and internal factors determining the degree to which the components of the product are coherent with one another. A product that is unusual but inappropriate will tend towards absurdity whilst one that is unoriginal but appropriate will tend towards *cliché* (Jackson & Messick, 1965).

Methodological approaches to evaluating creativity in musical composition have been developed which follow these theoretical accounts. Auh (2000), for example, describes her use of a scheme for the evaluation of creativity in human compositions which involves three assessment criteria: judgements of the originality or novelty of a piece; judgements of the appropriateness (the degree of tonal and rhythmic organisation and structure) of a piece; and judgements of expressiveness or musicality of the piece. Auh & Walker (1999) used this framework using five-point rating scales for the evaluation of the compositions of 18 school children by three expert, independent judges and obtained mean inter-judge correlations of $r = 0.71$. Auh & Johnston (2000) used the same framework for the evaluation of the compositions of 36 school children by three judges and obtained mean inter-judge correlations of $r = 0.88$. Regarding the

evaluation of appropriateness, Davidson & Welsh (1988) asked seven professional musicians to rank 10 melodies composed by conservatory students according to musical success. The resulting rank order of the melodies correlated significantly with the degree of musical experience of the student ($r = 0.86$).

Like many other theorists, Amabile (1996) proposes a conceptual definition of creativity in terms of processes which result in novel and appropriate solutions to heuristic, open-ended or ill-defined tasks (see also Simon, 1973). However, while agreeing that creativity can only be assessed through subjective assessments of products, Amabile criticises other approaches for using *a priori* theoretical definitions of creativity in their rating schemes and failing to distinguish creativity from other constructs. While a conceptual definition is important for guiding empirical research, a clear operational definition of creativity is necessary for the development of useful empirical methods of assessment. Accordingly, Amabile (1996) presents a consensual definition of creativity according to which a product is deemed creative to the extent that observers who are familiar with the relevant domain independently agree that it is creative. To the extent that this construct is valid in terms of internal consistency (independent judges agree in their subjective ratings of creativity) it will be possible to empirically examine the objective features or subjectively experienced dimensions of creative products that contribute to their perceived creativity.

Amabile (1996) has used this operational definition to develop an empirical methodology for evaluating creativity known as the *consensual assessment technique* (CAT). The essential features of this methodology are as follows. First, the task must be open ended enough to permit considerable flexibility and novelty in the response which must result in an observable product that can be rated by judges. Second, regarding the procedure, the judges should:

1. be experienced in the relevant domain;[3]
2. make independent assessments;
3. assess other aspects of the products such as technical accomplishment, aesthetic appeal or originality;
4. make relative judgements of each product in relation to the rest of the test items;
5. be presented with test items and provide ratings in orders randomised differently for each judge.

[3]Although, a number of studies have found high levels of agreement between judges with different levels of familiarity and experience (Amabile, 1996), this may be due to the technical simplicity of the tasks used.

Third, in terms of the analysis of the collected data, the most important issue is to determine the interjudge reliability of the subjective rating scales. Subject to high levels of reliability, creativity ratings may be correlated with other objective features or subjectively experienced dimensions of creative products.

A large number of experimental studies of verbal, artistic and problem solving creativity have demonstrated the ability of the CAT to obtain reliable subjective assessments of creativity in a range of domains (see Amabile, 1996, ch. 3, for a review). In recent years, the CAT has also been used successfully in assessing the musical compositions of students and school children (Brinkman, 1999; Hickey, 2001; Priest, 2001; Webster & Hickey, 1995). Brinkman (1999) used the CAT to examine relationships between creativity style (adaptor or innovator) and degrees of constraint in the problem specification. Three expert judges rated 64 melodies composed by high school students on seven-point scales of originality, craftsmanship and aesthetic value. The interjudge reliabilities obtained were 0.84 for originality, 0.77 for craftsmanship and 0.76 for aesthetic value for a combined total of 0.807. In research examining student's assessments of musical creativity in relation to their own ability to function creatively as composers, Priest (2001) used the CAT to place 54 university students into high-, middle- and low-creativity groups. Eight independent judges rated melodies composed by the students on continuous scales of creativity, melodic interest, rhythmic interest and personal preference yielding interjudge reliabilities of 0.81, 0.79, 0.85 and 0.84 respectively.

In an examination of a range of assessment scales for evaluating children's compositions, Webster & Hickey (1995) studied the relationship between ratings of craftsmanship, originality/creativity and aesthetic value obtained by the consensual assessment technique and other existing rating scales for music. Sub-items of these scales were categorised according to style (implicit or explicit) and content (global or specific). In a study in which four expert judges rated 10 children's compositions, Webster & Hickey (1995) found that implicitly defined rating scales tend to yield higher interjudge reliabilities (see also Amabile, 1996). However, while rating scales involving global and implicit definitions are better at predicting originality, creativity and aesthetic value, those involving explicit and specific definitions are better at predicting craftsmanship. In further research, Hickey (2000, 2001) asked which judges are appropriate in the context of evaluating children's compositions. Five groups of judges (music teachers, composers, music theorists, seventh-grade children and second-grade children) were asked to rate 12 pieces composed by fourth- and fifth-grade children on seven-point scales for creativity, craftsmanship and aesthetic ap-

peal. The results demonstrated that the most reliable judges were the music teachers and the least reliable the composers.

In summary, the CAT offers a methodological approach for evaluating computational models of musical compositions which is capable of overcoming the limitations of the Turing test approach discussed in §9.2.3. First, the methodology explicitly requires the use of appropriate judges; those with considerable practical experience and theoretical knowledge of the task domain. Second, since it has been developed for research on human creativity, no explicit mention is made of the computer generated origins of the artefacts; this should help avoid any potential biases due to a perception of the task as a challenge to catch out the computer. Third, and most importantly, the methodology allows the possibility of examining in more detail the objective and subjective dimensions of the generated products. Crucially, the objective attributes of the products may include features of the generative models (corresponding to cognitive or stylistic hypotheses) which produced them. In this way, it is possible to empirically compare different musicological theories of a given style or hypotheses about the cognitive processes involved in generating creative compositions in that style.

9.3 Experimental Hypotheses

Following Johnson-Laird (1991), the goal in this chapter is to examine the computational constraints of the task of composing a melody in two ways: first, to examine whether the trained finite context grammars developed in Chapters 6, 7 and 8 are capable of meeting the task demands of composing successful melodies or whether more expressive grammars are needed; and second, to examine which representational structures are necessary for the composition of successful melodies (see §9.2.1).

The experiment reported in this chapter was designed to test the hypothesis that the statistical models developed in Chapters 6 and 7 are capable of generating melodies which are deemed creative in the context of a specified stylistic tradition. To this end three multiple viewpoint systems trained on the chorale melodies in Dataset 2 (see Chapter 4) are used to generate melodies which are then empirically evaluated. The three systems in question are as follows: System A is the single viewpoint system (comprising `cpitch` alone) which was used in Chapter 6; System B is the multiple viewpoint system developed through feature selection in §8.7 to provide the closest fit to the experimental data of Manzara *et al.* (1992); and System C is the multiple viewpoint system

System	Viewpoints	H
A	`cpitch`	2.337[a]
B	`cpintfip`, `cpintfref`⊗`dur-ratio`, `thrfiph`	2.163
C	`cpint`⊗`dur`, `cpintfref`⊗`cpintfip`, `cpitch`⊗`dur` `cpintfref`⊗`fib`, `thrtactus`, `cpintfref`⊗`dur`, `cpint`⊗`dur-ratio`, `cpintfip`, `thrfiph`	1.953

[a]The discrepancy between this value and that shown in Table 6.7 is due to the fact that different parameterisations of the STM were used as explained in §7.4.

Table 9.1: The component viewpoints of multiple viewpoint systems A, B and C and their associated entropies computed by 10-fold cross-validation over Dataset 2.

developed through feature selection in §7.5.2 to yield the lowest model uncertainty over Dataset 2. Each system was parameterised optimally as discussed in Chapter 7 and differs only in the viewpoints it uses as shown in Table 9.1.

This work differs in a number of ways from previous approaches to the use of statistical models for generating music. One of the most salient omissions from our modelling strategy is that in contrast to the n-gram models of Hall & Smith (1996) and the neural network models of Hild *et al.* (1992), no symbolic constraints were imposed on the generation of compositions. Since the goal of the current research was to examine the synthetic capabilities of purely statistical, data-driven models of melodic structure, this approach was followed in order to focus the analysis more sharply on the inherent capacities of statistical finite context grammars.

In spite of this omission, the strategy employed improves on previous research in a number of ways. First, the variable order selection policy of PPM* is used to address concerns that low, fixed order models have a tendency to generate features uncharacteristic of the target style (Ponsford *et al.*, 1999). Also, as discussed in Chapter 6, other parameters of the models have been optimised to improve prediction performance over a range of different melodic styles. Second, in contrast to other approaches (*e.g.,* Triviño-Rodriguez & Morales-Bueno, 2001), Systems B and C operate over rich representational spaces supplied by the multiple viewpoint framework. In addition, and in contrast to the research of Conklin & Witten (1995), the viewpoints used in Systems B and C were selected on the basis of objective and empirical criteria. Also the systems use a novel model combination strategy, which was shown in Chapter 7 to improve prediction performance over Dataset 2.

Third, while the vast majority of previous approaches have used sequential random sampling to generate music from statistical models, in the present

research melodies were generated using Metropolis sampling (see §9.2.2). It is expected that this method will be capable of generating melodies which are more representative of the inherent capacities of the systems. It is worth emphasising that Metropolis sampling is not being proposed as a cognitive model of melodic composition but is used merely as a means of generating melodies which reflect the internal state of knowledge and capacities of the trained models.

Finally, in order to evaluate the systems as computational models of melodic composition, a methodology based on the CAT was developed and applied (see §9.2.4). The methodology, described fully in §9.4, involves the use of expert judges to obtain ratings of the stylistic success, originality and creativity of computer generated compositions and existing compositions in the target genre. It is hypothesised that one or more of the three systems shown in Table 9.1 will be capable of consistently generating compositions which are rated equally well on these scales as the chorale melodies in Dataset 2. The empirical nature of this methodology makes it preferable to the exclusively qualitative analyses which are typically adopted and, following the arguments made in §9.2.3 and §9.2.4, it is expected to yield more revealing results than the Turing test methodology used by Hall & Smith (1996) and Triviño-Rodriguez & Morales-Bueno (2001).

The purpose of using three different systems is to examine which representational structures are necessary to achieve competent generation of melodies. For each system, a null hypothesis is constructed according to which the system is capable of generating melodies which are rated as being equally successful, original and creative examples of a target style as existing, human composed melodies in that style. Assuming that the systems are unlikely to produce melodies that are more stylistically successful than existing melodies in the style, the null hypothesis for System A, which is something of a straw man, is expected to be rejected. In the case of Systems B and C, however, there are reasons to expect that the null hypothesis will be retained.

In the case of System B, the discussion of perceptual constraints on compositional grammars is relevant, especially the proposal of Baroni (1999) that composition and listening involve equivalent grammatical structures (see §9.2.1). If the representational structures underlying the perception and composition of music are very similar, we would expect grammars which model perceptual processes well to be able to generate satisfactory compositions. To the extent that System B represents a satisfactory model of the perception of pitch structure in the chorale genre, we may expect to retain the null hypothesis for this system.

In Chapter 8, a relationship was demonstrated between model uncertainty and fit to the human expectancy data obtained by Manzara *et al.* (1992) suggesting that human perceptual systems may base their predictions on representational features which reduce uncertainty. In terms of model selection for music generation, Conklin & Witten (1995) proposed that highly predictive theories of a musical style, as measured by entropy, will also be capable of generating original and acceptable works in the style. Table 9.1 shows that Systems A, B and C in turn exhibit decreasing uncertainty (cross entropy computed by 10-fold cross-validation) in predicting unseen melodies from Dataset 2. On this basis, it may be expected that the null hypothesis for System C will be retained.

9.4 Experimental Methodology

9.4.1 Judges

The judges used in this experiment were 16 members of the research and student communities of the music departments at: City University, London; Goldsmiths College, University of London; and the Royal College of Music, London. The ages of the judges, of whom five were male and eleven female, ranged between 20 and 46 years (mean 25.9, SD 6.5). They had been musically trained in a formal context for between 2 and 40 years (mean 13.8, SD 9.4) and all reported having moderate or high familiarity with the chorale genre. Specifically seven judges were highly familiar while nine were moderately familiar with the genre. Results were also collected and discarded from six judges who, in spite of having a musical background, reported having little or no familiarity with the genre. All judges received a nominal payment for participating in the experiment which lasted for approximately an hour.

9.4.2 Apparatus and Stimulus Materials

The stimuli consisted of 28 chorale melodies of which seven were selected from Dataset 2 (see Chapter 4) and seven each were generated by Systems A, B and C. The chorale melodies selected from Dataset 2 are notated in full in Appendix C. They were randomly selected from the set of chorales falling in the midrange of the distribution of cross entropy values computed using System A. All seven chorales are in common time; six are in a major key and one in a minor key (Chorale 238, BWV 310). The length of the seven melodies ranges from 8 to 14 bars (mean 11.14) and 33 to 57 events (mean 43.43).

Using each of the three multiple viewpoint systems, seven chorales were

generated using 5000 iterations of Metropolis sampling with the seven chorales selected from Dataset 2 as the initial states (see §9.2.2). In each case, only the pitches were sampled from the systems; the time and key signatures as well as rhythmic and phrase structure were taken from the chorale melody used as the initial state or *base chorale melody*. The chorale melodies generated by Systems A, B and C are notated in Appendices D, E and F respectively. The seven original chorale melodies selected were removed from the datasets used to train all systems.

In the interests of facilitating concise discussion of the individual test items, the following scheme will be used to refer to each melody. The seven original chorale melodies from Dataset 2 will be referred to in full by their number (*e.g.,* "Chorale 141") while the generated melodies will be referred to using a combination of the abbreviated name of the system and the base chorale employed in their generation. To give an example, "A141", "B141" and "C141" refer respectively to the melodies generated by Systems A, B and C with base chorale 141.

Each chorale melody was generated as a quantised MIDI file (Rothstein, 1992). A subtle pattern of velocity accents was added to emphasise the metric structure and a single crotchet rest was added after each fermata to emphasise the phrase structure. The stimuli were recorded to CD quality audio files on a PC computer using the standard piano tone of a Roland XP10 synthesiser connected via the MIDI interface of a Terratec EWS88 MT soundcard. All chorales were recorded at a uniform tempo of 90 beats per minute. The stimuli were presented to the judges over Technics RP-F290 stereo headphones fed from a laptop PC running a software media player. The judges recorded their responses in writing in a response booklet.

9.4.3 Procedure

The judges supplied their responses individually and received instructions in verbal and written forms. They were told that they would hear a series of chorale melodies in the style of Lutheran hymns. Their task was to listen to each melody in its entirety before answering four questions about the melody by placing circles at appropriate locations on discrete scales provided in the response booklet.[4] The first question was "How successful is the composition as a chorale melody?" Judges were advised that their answers should reflect

[4]One exception was made for a judge who was blind. In this case, the questions were read to the judge whose verbal responses were recorded by the experimenter. Since the judge listened to the stimuli on headphones and due to the randomised presentation order, the experimenter was unaware of which stimulus any given response referred to.

such factors as conformity to important stylistic features; tonal organisation; melodic shape and interval structure; and melodic form. The second question was "How original is the composition as a chorale melody?" Judges were advised that their answers should reflect the extent to which they felt that the composition is novel or original in the context of existing works in the style. The third question was "How creative is the composition as a chorale melody?" Judges were advised that their answers should reflect their own subjective definition of creativity. Answers to the first three questions were given on seven-point numerical scales, ranging from one to seven, with anchors marked low (one), medium (four) and high (seven). In an attempt to ensure an analytic approach to the task, judges were asked to briefly justify their responses to the first three questions. Following Webster & Hickey (1995), the ratings of stylistic success were presented in more specific terms than the ratings for originality and creativity. The final question was "Do you recognise the melody?" Judges were advised to answer in the affirmative only if they could specifically identify the composition as one they were already familiar with.

It was explained to judges that after all four questions had been answered for a given melody, they could listen to the next melody by pressing a single key on the computer keyboard. Judges were asked to bear in mind that their task was to rate the composition of each melody rather than the performance and were urged to use the full range of the seven-point scales, limiting ratings of 1 and 7 to extreme cases. There were no constraints placed on the time taken to answer the questions for each melody.

The experiment began with a practice session during which judges heard two melodies from the same genre (but not one of those in the test set). The two melodies chosen were Chorales 61 (BWV 159) and 151 (BWV 379) which were used in Chapter 8 (see Figure 8.6). These practice trials were intended to set a judgemental standard for the subsequent test session. This represents a departure from the CAT in which judges are encouraged to make their ratings of any given test item in relation to the others by experiencing all test items before proceeding to make their ratings. However, in this experiment it was intended that the judges should use their expertise to rate the test items in relation to an absolute standard represented by the body of existing chorale melodies. Judges responded as described above for both of the items in the practice block. The experimenter remained in the room for the duration of the practice session after which the judges were given an opportunity to ask any further questions. The experimenter then left the room before the start of the test session.

In the test session, the 28 melodies were presented to the judges who re-

sponded to the questions. The melodies were presented in random order subject to the constraints that no melody generated by the same system nor based on the same chorale should be presented sequentially. A reverse counterbalanced design was used for the test session with eight of the judges listening to the melodies presented in one such order and the other eight listening to them in the reverse order.

After completing the test session, the judges were asked to fill out a short questionnaire detailing their age, sex, number of years of music training (instrument and theory) and familiarity with the chorales harmonised by J. S. Bach (high/medium/low).

9.5 Results

9.5.1 Inter-judge Consistency

In the analysis of the results, the data collected for the two chorale melodies in the practice block were discarded. All analyses to be reported concern only the 28 chorale melodies from the main test session of the experimental trials. The first issue to examine concerns the consistency of the ratings across judges on the scales of success, originality and creativity. For the originality and creativity ratings only 58 and 24 of the 120 respective pairwise comparisons were significant at $p < 0.05$ with mean coefficients of $[r(26) = -0.021, p = 0.92]$ and $[r(26) = 0.027, p = 0.89]$ respectively. This lack of consistency in the ratings for originality and creativity may have been a result of the departure from the methodological conventions of the CAT in terms of encouraging the judges to consider the test items in relation to one another (see §9.4.3). Alternatively, it may reflect the views expressed by some judges that the concepts of originality and (especially) creativity did not make a great deal of sense to them in the context of this simple and traditional vocal style. In light of the lack of observed consistency, the originality and creativity ratings were not subjected to further analysis.

For the ratings of stylistic success, however, all but two of the 120 pairwise correlations between judges were significant at $p < 0.05$ with a mean coefficient of $[r(26) = 0.65, p < 0.001]$. Since there was no apparent reason to reject the judges involved in the two non-significant correlations, they were not excluded in subsequent analyses. The high levels of inter-judge consistency found for the success ratings warrant the averaging the ratings for each test item across individual judges for use in subsequent analyses.

Chorale	249	238	365	264	44	141	147
n	1	2	4	1	6	1	8

Table 9.2: The number of judges (n) who recognised each of the seven original chorale melodies in the test set.

9.5.2 Presentation Order and Prior Familiarity

One factor which might have influenced the ratings of the judges is the order in which they listened to the test items. However, the correlation between the mean success ratings for judges in the two groups was $r(26) = 0.91, p < 0.001$ indicating a high degree of consistency across the two orders of presentation and warranting the averaging of responses across the two groups of judges. Another factor which may have influenced the judges' ratings is their prior familiarity with the original chorale melodies used as test items. Of the 16 judges, nine reported recognising one or more of the original chorale melodies. Each of the seven original chorale melodies was recognised by at least one judge with a total of 23 recognised items over all 112 ratings obtained for the seven melodies from 16 judges (see Table 9.2). Interestingly, one of the nine judges who reported prior familiarity with one or more of the original chorale melodies also reported recognising three of the computer generated melodies (melodies A264, B264 and B44). These findings warrant treating the familiarity data with a certain amount of caution. Nonetheless, there was a tendency for the mean ratings of success to be slightly higher in cases where judges recognised the test item although a paired t test failed to reveal a significant difference [$t(6) = 2.07, p = 0.084$]. Since few of the original chorale melodies were familiar to more than a handful of the judges, it is hard to make a firm decision on the influence of prior familiarity and all ratings are included in subsequent analyses and no further attention is given to the issue.

9.5.3 Generative System and Base Chorale

The purpose of the next stage of analysis was to examine the influence of generative system on the ratings of stylistic success. The mean success ratings for each test item are shown in Table 9.3, with aggregate means for generative system and base chorale, and presented graphically in Figure 9.1. The mean ratings suggest that the original chorale melodies received higher success ratings than the computer generated melodies while the ratings for the latter show an influence of base chorale but not of generative system. Melody C249 is something of an exception receiving high average ratings of success. The

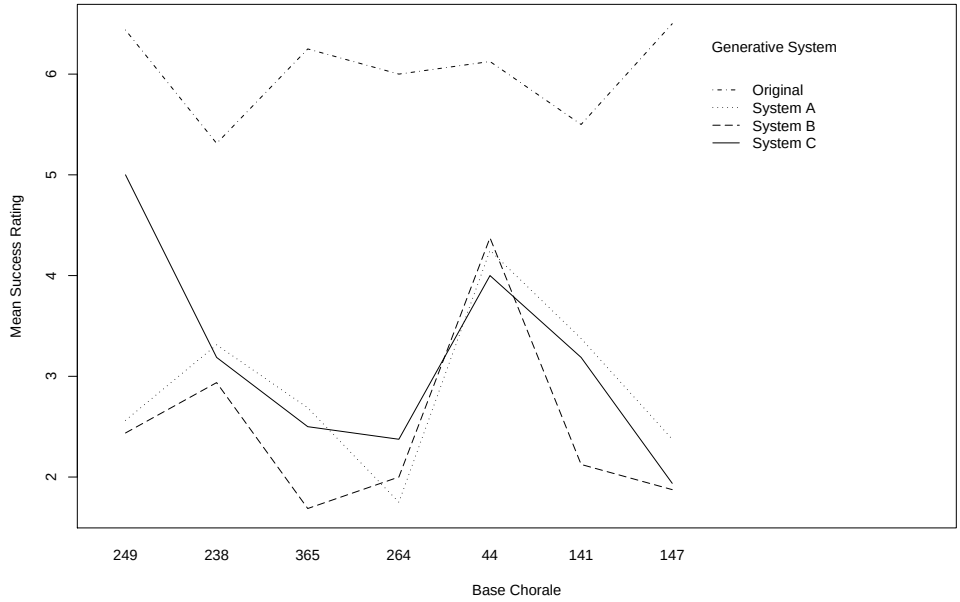


Figure 9.1: The mean success ratings for each test item.

planned analysis was a multivariate ANOVA using within-subjects factors for generative system with 4 levels (Original, System A, System B and System C) and base chorale with 7 levels (249, 238, 365, 264, 44, 153 and 147) with the null hypotheses of no main or interaction effects of generative system or base chorale. However, Levene's test revealed significant violations of the assumption of homogeneity of variance with respect to the factor for generative system $[F(3) = 6.58, p < 0.001]$. In light of this, Friedman rank sum tests were performed as a non-parametric alternative to the two-way ANOVA. Unfortunately this test does not allow the examination of interactions between the two factors.

The first analysis examined the influence of generative system in an unreplicated complete blocked design using the mean success ratings aggregated for each subject and generative system across the individual base chorales. Summary statistics for this data are shown in Table 9.4. The Friedman test revealed a significant within-subject effect of generative system on the mean success ratings $[\chi^2(3) = 33.4, p < 0.001]$. Pairwise comparisons of the factor levels were carried out using Wilcoxon rank sum tests with Holm's Bonferroni correction for multiple comparisons. The results indicate that the ratings for the original chorale melodies differ significantly from the ratings of melodies generated by all three computational systems ($p < 0.001$). Furthermore, the mean success ratings for the melodies generated by System B were found to be significantly

	System A	System B	System C	Original	Average
249	2.56	2.44	5.00	6.44	4.11
238	3.31	2.94	3.19	5.31	3.69
365	2.69	1.69	2.50	6.25	3.28
264	1.75	2.00	2.38	6.00	3.03
44	4.25	4.38	4.00	6.12	4.69
141	3.38	2.12	3.19	5.50	3.55
147	2.38	1.88	1.94	6.50	3.17
Average	2.90	2.49	3.17	6.02	3.65

Table 9.3: The mean success ratings for each test item and means aggregated by generative system and base chorale.

different from those of the melodies generated by Systems A and C ($p = 0.027$). These results suggest that none of the systems are capable of consistently generating chorale melodies which are rated as equally successful stylistically as those in Dataset 2 and that System B performed especially poorly.

The second analysis examined the influence of base chorale in an unreplicated complete blocked design using the mean success ratings aggregated for each subject and base chorale across the individual levels of generative system. Summary statistics for this data are shown in Table 9.5. The Friedman test revealed a significant within-subject effect of base chorale on the mean success ratings $[\chi^2(6) = 49.87, p < 0.001]$. Pairwise comparisons of the factor levels were carried out using Wilcoxon rank sum tests with Holm's Bonferroni correction for multiple comparisons. The results indicate that the mean rating for melodies generated with base chorale 44 was significantly different (at $p < 0.01$) from that of melodies generated with all other base chorales (except Chorale 249, $p = 0.072$). In addition, the mean rating of melodies generated with base chorale 249 was significantly different (at $p < 0.05$) from that of melodies generated with all other base chorales (except Chorales 44, see above, and 238, $p = 0.092$). An examination of the mean success ratings plot-

Statistic	System A	System B	System C	Original
Median	2.86	2.57	3.07	5.93
Q1	2.68	2.25	2.68	5.86
Q3	3.29	2.75	3.61	6.29
IQR	0.61	0.50	0.93	0.43

Table 9.4: The median, quartiles and inter-quartile range of the mean success ratings for each generative system.

Statistic	249	238	365	264	44	141	147
Median	4.12	3.75	3.25	3.00	4.75	3.75	3.25
Q1	4.00	3.44	2.75	2.75	4.44	3.19	2.94
Q2	4.25	4.00	3.81	3.31	5.06	3.75	3.31
IQR	0.25	0.56	1.06	0.56	0.62	0.56	0.37

Table 9.5: The median, quartiles and inter-quartile range of the mean success ratings for each base chorale.

ted in Figure 9.1, suggests that the significant effects of base chorale are largely due to the relatively high mean success rating of melody C249 and all computer generated melodies with Chorale 44 as their base melody.

9.5.4 Objective Features of the Chorales

A question which arises from the previous analyses concerns the objective musical features of the test items which were used by the judges in making their ratings of stylistic success. Given the finding that the generated melodies were, in general, rated as less successful than the original chorale melodies, an answer to this question could shed light on how the systems are lacking as models of composition. In order to address this issue, a simple qualitative analysis of the test items was carried out and used to develop a set of objective descriptors. These predictors were then applied in a series of multiple regression analyses using the three rating schemes, averaged across test items, as dependent variables. The descriptive variables and their quantitative coding are presented before discussing the results of the analyses.

The chorale melodies contained in Dataset 2 represent congregational hymns of the Lutheran church which were either composed for this purpose or adapted from existing pre-Reformation hymns and secular folk songs in the sixteenth and seventeenth centuries. Specifically, Dataset 2 contains a subset of the chorale melodies placed in the soprano voice and harmonised in four parts by J. S. Bach in the first half of the eighteenth century. Although they do reflect these diverse origins, these melodies are characterised by stepwise patterns of conjunct intervallic motion as well as simple, uniform rhythmic and metric structure. Phrase structure in the chorales is consistently notated by means of fermatas which emphasise both melodic sequences and the implied harmonic movement. The majority of phrases begin on the tonic, mediant or dominant scale degrees and end on the tonic or dominant with a cadence to the tonic virtually ubiquitous in the case of the final phrase.

The chorales generated by the three statistical systems are on the whole not

very stylistically characteristic of Dataset 2 although some are moderately successful. The melody generated by System C with base chorale 249, for example, is reasonably coherent in terms of melodic form although it lacks harmonic direction especially in the final phrase. The qualitative comments supplied by the judges to justify their ratings were used to identify a number of stylistic constraints describing the test items and distinguishing the original chorale melodies. These may be grouped into five general categories: pitch range; melodic structure; tonal structure; phrase structure; and rhythmic structure. The predictor variables presented below were developed to cover all five of these categories. The categories are very general and, along with some of the specific predictors, bear a certain resemblance to the rating scales of Cantometric analysis used in comparative ethnomusicology (Lomax, 1962).

Pitch Range The chorale melodies in Dataset 2 are written for the soprano voice and span a pitch range of approximately an octave above and below C_5 tending to favour the centre of this range. While the generated melodies are constrained to operate within this range, some tend towards unusually high or low tessitura. Examples of the former include the melodies B365 and B264 and of the latter the melody C238. The predictor variable *pitch centre* was developed to capture such intuitions. Following von Hippel (2000b), it is a quantitative variable reflecting the absolute distance, in semitones, of the mean pitch of a given melody from the mean pitch of Dataset 2 (approximately $B\flat_4$). Another issue to consider concerns the overall pitch range of the generated chorales. The chorale melodies in Dataset 2 span, on average, a range of just under an octave (mean pitch range = 11.8 semitones). By contrast, several of the 21 generated melodies span pitch ranges of 16 or 17 semitones (*e.g.,* melodies C147, B249 and A264) with a mean pitch range of 13.9 semitones. Others, such as melody A141 operate within a rather narrow range of pitch height. These qualitative considerations were captured in a quantitative predictor variable *pitch range* which represents the absolute distance, in semitones, of the pitch range of a given melody from the mean pitch range of Dataset 2.

Melodic Structure The generated melodies fail to consistently reproduce the salient melodic features of the chorales in Dataset 2 in a number of ways. Perhaps the most obvious of these is a common failure to maintain a stepwise pattern of conjunct movement. While some of the generated melodies, such as melodies C44 and C249, are relatively coherent, others, such as melody C147, contain stylistically uncharacteristic intervallic leaps of an octave or more. Of

9042 melodic intervals in the chorale melodies in Dataset 2, just 57 are greater than a perfect fifth and none exceeds an octave. In order to capture these deviations from the conventions of Dataset 2, a quantitative predictor variable called *interval size* was created which represents the number of intervals in a melody which are greater in size than a perfect octave. Apart from interval size, the generated chorales contain uncharacteristic discords such as the tritone in the penultimate bar of the melody B147 or the sevenths in the second phrase of the melody B238 and the third and final phrases of melody C147. Only 8 of the 9042 intervals in Dataset 2 are tritones or sevenths (or their enharmonic equivalents). In order to capture these deviations from the conventions of Dataset 2, a quantitative predictor variable called *interval dissonance* was created which represents the number of dissonant intervals greater than a perfect fourth in a given melody.[5]

Tonal Structure Since it operates exclusively over representations of pitch height, it is not surprising that most of the melodies generated by System A fail to establish a key note and exhibit little tonal structure. However, we might expect System B and, especially, System C to fare somewhat better in this regard.

While the comments of the judges suggest that this is not the case, it is quite possible that the judges arrived at a tonal interpretation at odds with the intended key of the base chorale. In order to independently estimate the perceived tonality of the melodies presented to the judges, a key finding algorithm was applied to each of the test items with the results shown in Table 9.6. The algorithm in question is the Krumhansl-Schmuckler key finding algorithm (Krumhansl, 1990) modified to use the revised key profiles developed by Temperley (1999).[6] Note that the algorithm assigns the correct keys to all seven of the original chorale melodies. While the suggested keys of the melodies generated by System A reflect the fact that it does not consider tonal constraints, the melodies generated by Systems B and C retain the key of their base chorale in two and five cases respectively. Furthermore, especially in the case of the melodies generated by System C, deviations from the key of the base chorale tend to be to related keys (either in the circle of fifths or through relative and parallel major/minor relationships). This suggests a degree of success on the part of the more sophisticated systems in retaining the tonal characteristics of

[5]It should be noted that dissonance is used here in its musical sense; this predictor does not reflect an attempt to model sensory dissonance.

[6]While Temperley (1999) proposes a number of improvements to the Krumhansl-Schmuckler algorithm, the algorithm used here differs from the original only its use of the revised key profiles. Temperley derived the revised profiles through both trial and error and theoretical reasoning to address a number of concerns with the original profiles.

	System A	System B	System C	Original
249	F Major	G Minor	G Major	G Major
238	G Major	D Major	E Major	E Minor
365	B♭ Major	F♯ Minor	A Major	A Major
264	A Minor	B♭ Major	B♭ Major	B♭ Major
44	C Major	G Major	D Major	D Major
141	C Major	E Minor	A Major	A Major
147	G Major	E♭ Major	B♭ Major	E♭ Major

Table 9.6: The key returned by the key-finding algorithm of Temperley (1999) for each test item.

the base chorales.

Nonetheless, the generated melodies often exhibit a degree of chromaticism which is unacceptable in the style since it obscures the prevailing tonality. In many cases, this seems to be a result of the failure to represent mode in the multiple viewpoint systems. Chorale 238, for example, is written in the key of E minor but the melody C238 contains accidentals from the key of E major (C♯, G♯ and D♯). Similar comments apply to the melodies C264, C44 and B44. Other generated melodies appear to include chromatic elements from related keys as in the case of the melody C141, intended to be in the key of A major, which includes the accidentals D♯ and A♯ from the related keys of E and B major. On the basis of these considerations, a quantitative predictor called *chromaticism* was developed which represents the number of tones which are chromatic in the key suggested by the key-finding algorithm with the assumption that this reflects the tonality induced by the judges in listening to the melodies.

Phrase Structure The generated chorales also typically fail to reproduce the simple implied harmonic rhythm of the original chorale melodies and its characteristically strong relationship to phrase structure. In particular, while some of the generated melodies close on the tonic, (*e.g.,* melodies C249, C264 and C141), many fail to imply harmonic closure in a stylistically satisfactory manner (*e.g.,* melodies C238, C365, C44 and C147). The generated melody C44, for example, breaks the implied I-V-I harmonic movement of Chorale 44, especially in the final phrase. In order to capture such effects, a dummy variable called *harmonic closure* was created which assumes a value of zero if a melody closes on the tonic of the key assigned by the key-finding algorithm described above and one otherwise. In addition, the generated melodies frequently fail to respect thematic repetition and development of melodic material often embedded within the phrase structure of the chorales. To take an obvious example,

the two opening phrases of Chorale 264 are repeated exactly in the fifth phrase. However, the structural coherence this brings to the melody is almost entirely obliterated in the generated melody C264. More subtle examples of failures to repeat and transform melodic motifs abound in the generated melodies. However, these kinds of repetition and development of melodic material within the phrase structure of a chorale are difficult to quantify and are not represented in the present model. Instead, as an extremely simple indicator of complexity in phrase structure, a second dummy variable *phrase length* was created which assumes a value of zero if all phrases are of equal length (in terms of tactus beats) and one otherwise.

Rhythmic Structure Although the chorale melodies in Dataset 2 tend to be very simple rhythmically, the finding of significant (and marginally non-significant) effects of base chorale in the previous analyses raises the question of whether rhythmic structure may have influenced the ratings of the judges. Furthermore, the comments of some judges revealed that they were taking considerable account of rhythmic structure in making their judgements. For these reasons, three further quantitative predictors modelling rhythmic features were adapted from the expectancy-based model of melodic complexity developed by Eerola & North (2000). First, *rhythmic density* is a quantitative predictor representing the mean number of events per tactus beat. Second, *rhythmic variability* is a quantitative predictor which models the degree of changes in note duration and is coded as the standard deviation of the log of the event durations in a melody. Finally, *syncopation* represents the degree of syncopation by assigning tones a pulse strength within a metric hierarchy (Lerdahl & Jackendoff, 1983; Palmer & Krumhansl, 1990) and taking the average strengths of all the tones in a given melody. Pulses are coded such that lower values are assigned to tones on metrically stronger beats. As noted by Eerola & North (2000), all three quantities have been demonstrated to increase the difficulty of perceiving or producing melodies (Clarke, 1985; Conley, 1981; Povel & Essens, 1985).

The comments of the judges in justifying their judgements of the success of the melodies generally reflected the considerations involved in the development of these predictors. They frequently commented that the incoherence and complexity of the generated melodies in terms of pitch range, form, tonality and melodic structure would make them very difficult to sing or to harmonise.

The mean ratings of success for each test item were regressed on the ten predictor variables introduced above in a multiple regression analysis. Of the

Predictor	β	Std. Error	t	p
Pitch Range	-0.2854	0.0799	-3.57	< 0.01
Pitch Centre	-0.2066	0.1026	-2.01	< 0.1
Interval Dissonance	-0.7047	0.2776	-2.54	< 0.05
Chromaticism	-0.2716	0.0336	-8.09	< 0.001
Phrase Length	-0.5258	0.2759	-1.91	< 0.1
Overall model: $R = 0.922$, $R^2_{adj} = 0.817$, $F(5, 22) = 25.04$, $p < 0.001$				

Table 9.7: Multiple regression results for the mean success ratings of each test melody.

pairwise correlations between the predictors, the following were significant at $p < 0.05$: interval size correlated positively with interval dissonance ($r = 0.6$) and chromaticism ($r = 0.39$); harmonic closure correlated positively with chromaticism ($r = 0.49$); rhythmic variation correlated positively with syncopation ($r = 0.61$) and phrase length ($r = 0.73$); and rhythmic density correlated positively with syncopation ($r = 0.62$) and negatively with phrase length ($r = -0.54$). Given this collinearity, redundant predictors were removed from the regression model through backward stepwise elimination using the Akaike Information Criterion (AIC). For a regression model with p predictors and n observations:

$$AIC = n \log(RSS/n) + 2p + c$$

where c is a constant and RSS is the residual sum of squares of the model (Venables & Ripley, 2002). Since larger models will provide better fits, this criterion attempts to balance model size, represented by p, with the fit of the model to the dependent variable, represented by RSS.

Regarding the predictors, more positive values indicate greater deviance from the standards of Dataset 2 (for pitch range and centre) or increased melodic complexity (for the remaining predictors). On this basis, it is expected that each predictor will exhibit a negative relationship with the success ratings. The results of the multiple regression analysis with the mean success ratings as the dependent variable are shown in Table 9.7. The overall model accounts for a significant proportion (approximately 82%) of the variance in the mean success ratings. Apart from rhythmic structure, at least one predictor from each of the five categories made a significant (or marginally significant) contribution to the fit of the model. Furthermore, the coefficients for all of the selected predictors are negative as predicted. Overall, the model indicates that the judged success of a test item decreases as its pitch range and centre depart from the mean range and centre of Dataset 2, with increasing numbers of dissonant in-

Stage	Viewpoint Added	H
1	cpint⊗dur	2.214
2	cpintfref⊗mode	2.006
3	cpintfref⊗cpintfip	1.961
4	cpitch⊗dur	1.943
5	thrfiph	1.933
6	cpintfref⊗liph	1.925
7	cpint⊗dur-ratio	1.919
8	cpint⊗inscale	1.917
9	cpintfref⊗dur	1.912
10	cpintfiph	1.911

Table 9.8: The results of viewpoint selection for reduced entropy over Dataset 2 using an extended feature set.

tervals and chromatic tones and if it has unequal phrase lengths.

9.5.5 Improving the Computational Systems

The analysis of the generated chorales conducted in §9.5.4 suggests that several important stylistic constraints of the chorale genre are lacking in the computational systems examined. These constraints primarily concern pitch range, intervallic structure and tonal structure. In order to examine whether the systems can be improved to respect such constraints, a number of viewpoints were added to those used in selecting System C and the resulting models were analysed in the context of prediction performance. Regarding tonal structure, it seems likely that the evident confusion of parallel minor and major modes is due to the failure of any of the systems to represent mode. In order to examine this hypothesis, a linked viewpoint cpintfref⊗mode was included in the extended feature space. Furthermore, it is hypothesised that the skewed distribution of pitch classes at phrase beginnings and endings can be more adequately modelled by two linked viewpoints cpintfref⊗fiph and cpintfref⊗liph. On the hypothesis that intervallic structure is constrained by tonal structure, another linked viewpoint cpint⊗inscale was also included. Finally, in an effort to represent potential constraints on pitch range and centre, a new viewpoint tessitura was created which assumes a value of 0 if the pitch of an event is within 1 standard deviation of the mean pitch of Dataset 2, -1 if it is below this range and 1 if it is above. The linked viewpoint tessitura⊗cpint was used in the feature set to represent the potentially complementary constraints of pitch height and interval size and direction.

These five viewpoints were added to the set, shown in Tables 5.2 and 5.4,

used in the feature selection experiment of Chapter 7 which led to the development of System C. The feature selection algorithm discussed in §7.4 was run over this extended feature space with the empty multiple viewpoint system as its start state to select feature subsets which reduce model uncertainty. The results of feature selection are shown in Table 9.8. In general, the resulting multiple viewpoint system (referred to as System D) shows a great deal of overlap with System C. Just three of the nine viewpoints present in System C were not selected for inclusion in System D: `cpintfref`⊗`fib`; `thrtactus`; and `cpintfip`. It seems likely that this is due to fact that three of the five new viewpoints were selected for inclusion in System D: `cpintfref`⊗`mode`; `cpintfref`⊗`liph`; and `cpint`⊗`inscale`. The first and second of these viewpoints, in particular, were added early in the selection process. In addition, the existing viewpoint `cpintfiph` was added in the final stage of feature selection. Finally, it is important to note that System D exhibits a lower average entropy ($H = 1.911$) than System C ($H = 1.953$) in predicting unseen compositions in Dataset 2. The significance of this difference was confirmed by paired t tests over all 185 chorale melodies [$t(184) = 5.985, p < 0.001$] and averaged for each 10-fold partition of the dataset [$t(9) = 12.008, p < 0.001$] (see §7.5.1).

9.6 Discussion and Conclusions

The goal of these experiments was to examine the intrinsic computational-level demands of the task of melodic composition. In particular, the aim was to examine constraints placed on the representational primitives and expressive power of the computational system in the composition of a successful melody. In order to achieve these goals, three multiple viewpoint systems were developed: first, a simple system which represents only pitch height; second, a system which provides a close fit to the expectations of listeners to chorale melodies; and third, a system which exhibits relatively low uncertainty in predicting events in unseen chorale melodies. Within the context of these three systems, finite context grammars were trained on Dataset 2 and used to generate new pitches for seven of the chorales in that corpus. Musically trained judges who were familiar with the domain were asked to rate the original and generated melodies in terms of their perceived success, originality and creativity as chorale melodies.

For each system a null hypothesis was constructed according to which the system would be capable of generating melodies which are rated as being equally successful, original and creative examples of the style as the original melodies. The originality and creativity ratings showed little consistency across

judges and were discarded. Regarding the ratings for success, however, the results of the analysis suggested significant effects of generative system and base chorale. Further analyses indicated that the effects were attributable to the fact that the original chorale melodies were rated as more stylistically successful than the computer generated melodies. On this basis, the null hypotheses may be rejected for all three systems; none of the computational systems is capable of consistently generating chorale melodies which are rated as equally successful examples of the style as the original chorale melodies in Dataset 2. In a second analysis, the qualitative comments of the judges were used to derive a number of predictors describing objective features of the test items. The results of a multiple regression analysis demonstrated that the success ratings tended to decrease when the pitch range and pitch centre diverged from those of Dataset 2, with increasing numbers of dissonant intervals and chromatic tones, and with uneven phrase lengths.

This analysis of the relationship between objective features of the chorales and the ratings of stylistic success suggested some ways in which the models could be improved to better reflect the constraints of the style. Several viewpoints were developed in an effort to represent potential constraints on tonal-harmonic structure, intervallic structure and pitch range. In a subsequent feature selection experiment, three of these new viewpoints were selected resulting in System D which has significant overlap with System C but which exhibits significantly lower uncertainty in predicting unseen chorale melodies in Dataset 2. Appendix G presents a preliminary investigation into the capacity of System D to generate stylistically successful chorale melodies.

Some discussion is warranted of the finding that the statistical finite context grammars used in the current research failed to match the computational demands of the task of composing chorale melodies regardless of the representational primitives used. Since steps were taken to address the limitations of previous context modelling approaches to generating music, it might be concluded that more powerful grammars are required to successfully achieve this task. The question of how powerful the grammar needs to be is an empirical one which should be addressed in future research. In this regard, a number of approaches can be envisaged. First, it is possible that a further analysis of the capacities of finite context modelling systems will prove fruitful. Future research should use the methodology developed here to analyse System D, identify its weaknesses and elaborate it further. Second, it remains possible that the MCMC sampling procedure was partly responsible for the negative result, in spite of the fact that this method represents an improvement (in terms of obtaining an

unbiased sample from the target distribution) over the sequential random sampling method used in previous research. More structured generation strategies, such as pattern based sampling techniques (Conklin, 2003; Hörnel, 1997), may be capable of conserving phrase level regularities and repetitions in a way that the models examined here clearly were not. Third, symbolic constraints may be employed to examine hypotheses about the nature of compositional competence within the framework of finite context modelling (Hall & Smith, 1996; Hild *et al.*, 1992; Povel, 2004). An approach such as this might prove capable of providing more satisfactory models of intra-opus regularities than the short-term models used here. Finally, future developments in neural network research (see §3.5) may lead to architectures and training strategies which allow networks to acquire representations of constraints of a sufficient expressive power to successfully model the cognitive process of melody composition.

A number of issues concerning the methodological approach also warrant discussion. Perhaps most significantly, the adapted CAT yielded insightful results for ratings of stylistic success in spite of the fact that the judges were encouraged to rate the test items according to an absolute standard (*cf.* Amabile, 1996). However, the results highlight a number of recommendations for future research. First, future research should completely avoid the possibility of method artefacts by randomising the presentation order of both test items and practice items for each judge and also randomising the order in which each rating scale is presented (Amabile, 1996). Second, the comments of the judges sometimes reflected the influence of aesthetic appeal on their judgements (*e.g.,* "doesn't work . . . but endearing and engaging"). In the interests of delineating subjective dimensions of the product domain in the assessment task (Amabile, 1996), judges should also be asked to rate the test items on aesthetic appeal. Third, although the influences of prior familiarity with the test items were ambiguous, efforts should be made to avoid any potential bias resulting from recognition of the stimuli. Finally, future work should examine why the inter-judge reliability was so low for the originality and creativity ratings. Possible causes for this finding include the fact that judges were not encouraged to make relative assessments of the test items (see §9.4.3) or the degree to which the concepts of originality and creativity apply to this simple and traditional vocal style. In any case, the present results suggest that the task of composing a stylistically successful chorale melody (regardless of its originality or creativity) presents significant challenges as a first step in modelling cognitive processes in composition.

Nonetheless, the methodological approach to evaluation derived from the

consensual assessment technique proved to be highly fruitful in examining the generated melodies in the context of existing pieces in the style. This methodology facilitated the empirical examination of specific hypotheses concerning the models through a detailed comparison of the generated and original melodies on a number of dimensions. It also permitted the examination of objective features of the melodies which influenced the ratings and the subsequent identification of weaknesses in the generative systems and directions for improving them. This provides a practical demonstration of the utility of analysis by synthesis in the context of modelling cognitive processes in composition as long as it is combined with an empirical methodology for evaluation such as the one developed here.

9.7 Summary

The goal in this chapter was to examine, at the computational level, the intrinsic demands of the task of composing a successful melody. Of particular interest were constraints placed on the degree of expressive power and the representational primitives of the compositional system. In order to achieve these goals, three multiple viewpoint systems developed in previous chapters were used to generate new pitches for seven of the chorale melodies in Dataset 2. In §9.3, null hypotheses were presented which stated that each of the three models would be capable of consistently generating chorale melodies which are rated as equally successful, original and creative examples of the style as the chorale melodies in Dataset 2. In order to examine these hypotheses experienced judges were asked to rate the generated melodies together with seven original chorale melodies on each of these three dimensions. The results, presented in §9.5, warrant the rejection of the null hypothesis for all three of the systems, mainly on the basis of the success ratings. In spite of steps taken to address some notable limitations of previous context modelling approaches to generating music, the finite context grammars making up these systems exhibited little ability to meet the computational demands of the task regardless of the representational primitives used. Nonetheless, a further analysis identified some objective features of the chorale melodies which exhibit significant relationships with the ratings of stylistic success. These results, in turn, suggested ways in which the computational models were failing to meet intrinsic stylistic constraints of the chorale genre. Adding certain viewpoints to the multiple viewpoint systems to address these concerns resulted in significant improvements in the prediction performance of the models. In contrast to previous approaches to the evaluation of

computational models of compositional ability, the methodological framework developed in this chapter enabled a detailed and empirical examination and comparison of melodies generated by a number of models, the identification of weaknesses of those models and their subsequent improvement.

CHAPTER 10

CONCLUSIONS

10.1 Dissertation Review

In Chapter 1, the present research was motivated in terms of a discrepancy between the development of sophisticated statistical models of musical structure in AI research and the predominant use of symbolic rule-based systems derived from expert music-theoretic knowledge in research on music perception. The former offer an opportunity to address the music-theoretic biases, inflexibility and cross-cultural limitations of the latter. The specific objectives of the present research were to develop powerful statistical models of melodic structure; to apply these models in the examination of specific hypotheses regarding cognitive processing in melody perception and composition; and to investigate and adopt appropriate methodologies for the empirical evaluation of such hypotheses. The principal claim investigated was that statistical models which acquire knowledge through the induction of regularities in existing music can, if examined with appropriate methodologies, provide significant insights into the cognitive processing involved in music perception and composition.

The methodological foundations for achieving the research objectives were presented in Chapter 2. In particular, arguments were presented for examining music perception and composition at the computational level, for following a machine learning approach and for evaluating cognitive models using empirical experiments. In Chapter 3, the finite context models used in this research were introduced in terms of the languages they can generate, their assumptions and the methodological constraints they impose. While these models suffer from

their limited expressive power, they are compatible with a machine learning approach unlike many more powerful classes of grammar. Several approaches to addressing the inherent limitations of these models were discussed in a review of their use in previous research for modelling musical structure, generating music and modelling music perception.

The corpora of melodic music used in the present research were introduced in Chapter 4 including Dataset 2 which consists of 185 of the chorale melodies harmonised by J. S. Bach and was used extensively in later chapters. Chapter 5 presented the scheme used to represent musical objects which takes as its musical surface the properties of discrete musical events at the note level. Events are composed of attributes representing properties related to event timing and pitch as well as other metric, tonal and phrase level features notated in the score. In the interests of endowing the representation scheme with greater structural generality, a multiple viewpoint framework (Conklin & Witten, 1995) was developed which permits the flexible representation of many different features derived from the musical surface. On the basis of previous research on music perception and computational music analysis, a collection of viewpoints was constructed to allow the representation of the pitch, rhythmic, tonal, metric and phrase structure of a melody as well as relationships between these structural domains.

In Chapter 6, a number of strategies for improving the performance of finite context models were examined empirically in the context of predicting the pitches of events in unseen melodies in a range of different styles. Some of these strategies concern the smoothing mechanism employed: first, several different techniques for computing the escape probabilities were examined; second, backoff smoothing was compared with interpolated smoothing; and third, a technique for removing the global order bound altogether was examined. Another technique examined, update exclusion, is an alternative policy for maintaining frequency counts. The final strategy examined combines the predictions of a trained model with those of a short-term model trained online during prediction of the current melody.

In a series of experiments, these techniques were applied incrementally to eight melodic datasets using cross entropy computed by 10-fold cross-validation on each dataset as the performance metric. The results demonstrated the consistent and significant performance improvements afforded by the use of escape method C (and AX with the short-term model), unbounded orders, interpolated smoothing and combining long- and short-term models. Furthermore, since these findings were obtained over a range of musical genres and generally cor-

roborate findings in data compression and statistical language modelling, there is reason to believe that the improvements afforded are robust across domains of application.

In Chapter 7, these improved models were applied within the context of the multiple viewpoint framework. A central issue when predicting melodies with multiple viewpoints concerns the method by which distributions estimated by different models are combined. In an experiment which compared the performance of a novel combination technique based on a weighted geometric mean with that of an existing technique based on a weighted arithmetic mean, the former was found to outperform the latter. This effect was much more pronounced when combining viewpoint models than the long- and short-term models. It was proposed that this asymmetry results from the fact that the former involves combining estimates derived from distinct data representations. In a second experiment, a feature selection algorithm was applied to select multiple viewpoint systems that are associated with lower cross entropy over Dataset 2. The final selected system is dominated by linked and threaded viewpoints which emphasise stylistic regularities in the corpus in terms of relative pitch structure, relationships between pitch and rhythmic structure and the influence of metric and phrase level salience.

The goal in Chapter 8 was to examine the cognitive processing involved in the generation of perceptual expectations while listening to a melody. The implication-realisation theory of Narmour (1990) is a detailed account of expectancy in melody according to which the expectations of a listener depend to a large extent on a small number of Gestalt-like rules which are held to be innate and universal. An alternative theory was presented which claims that observed patterns of melodic expectation can be accounted for in terms of the induction of statistical regularities existing in the music to which listeners are exposed.

In order to test the theory, three experiments were conducted to examine the correspondence between the patterns of expectation exhibited by listeners and those exhibited by the statistical models developed in Chapters 6 and 7 in the context of increasingly complex melodic stimuli. The question of which melodic features afford regularities capable of supporting the acquisition of the observed patterns of expectation was also addressed by selecting multiple viewpoint systems exhibiting closer fits to the behavioural data. The results demonstrate that the statistical models can account for the expectations of listeners as well as, or better, than the IR model especially when expectations were elicited in longer melodic contexts. The results also indicate that the statistical

model largely subsumes the function of the principles of Proximity and Reversal (Schellenberg, 1997) in accounting for the expectations of listeners, rendering the inclusion of these rules in an additional system of innate bottom-up predispositions unnecessary. Overall, the viewpoints selected in the experiments reflected a strong influence of interval structure, relative pitch structure and a relationship between these dimensions of pitch structure and rhythmic structure.

The goal in Chapter 9 was to examine the intrinsic demands of the task of composing a successful melody. Of particular interest were constraints placed on the degree of expressive power and the representational primitives of the compositional system. In order to achieve these goals, three multiple viewpoint systems developed in previous chapters were used to generate new pitches for seven of the chorale melodies in Dataset 2 using Metropolis sampling in place of the sequential random sampling method typically used. The null hypothesis stated that these systems are capable of consistently generating chorale melodies which would be rated as equally successful, original and creative examples of the style as the chorale melodies in Dataset 2. An adapted form of the Consensual Assessment Technique (Amabile, 1996) for the assessment of psychological components of human creativity was used to examine this hypothesis. In this methodology, experienced judges are asked to rate generated melodies together with original melodies on a number of dimensions.

In spite of steps taken to address some notable limitations of previous context modelling approaches to generating music, the results demonstrate that the finite context grammars making up these systems possess little ability to meet the computational demands of the task regardless of the representational primitives used. However, in contrast to previous approaches to the evaluation of computational models of compositional ability, the methodological approach enabled a further quantitative analysis of specific ways in which the computational models failed to represent some important stylistic constraints of the chorale genre. These failures were addressed by augmenting the multiple viewpoint systems with additional viewpoints resulting in significant improvements in prediction performance.

10.2 Research Contributions

In §2.3, a distinction was made between three different branches of AI each with its own motivations, goals and methodologies: basic AI; cognitive science; and applied AI. The present research makes direct contributions in the fields of

basic AI and, especially, cognitive science and indirectly contributes to the field of applied AI.

The goal of basic AI is to examine computational techniques which have the potential for simulating intelligent behaviour. Chapters 6 and 7 present an examination of the potential of a range of computational modelling techniques to simulate intelligent behaviour in the induction of regularities in existing corpora of melodic music and the use of these regularities in predicting unseen melodies. The techniques examined and the methodologies used to evaluate these techniques were drawn from the fields of data compression, statistical language modelling and machine learning. In empirically identifying a number of techniques which consistently improve the performance of finite context models of melodic music, the present research contributes to our basic understanding of computational models of intelligent behaviour in the induction and prediction of musical structure. In particular, Chapter 7 introduced a new technique based on a weighted geometric mean for combining the predictions of multiple models trained on different viewpoints which was shown to outperform an existing technique based on a weighted arithmetic mean.

Another contribution made in the present research was to use a feature selection algorithm to construct multiple viewpoint systems (see 5.2.3) on the basis of objective criteria rather than hand-crafting them on the basis of expert human knowledge as has been done in previous research (Conklin, 1990; Conklin & Witten, 1995). This has allowed the empirical examination of hypotheses regarding the degree to which different representational dimensions of a melody afford regularities that can be exploited by statistical models of melodic structure and melody perception.

The goal of cognitive-scientific research is to further our understanding of human cognition using computational techniques. Contributions to cognitive science were made in Chapters 8 and 9 where the statistical finite context models developed in Chapters 6 and 7 were used to examine computational level constraints on the cognitive tasks of perceiving melodic structure and composing melodies respectively. Specifically, the research reported in Chapter 8 proposed a theory of melodic expectancy based on statistical learning and adopted methodologies from cognitive science and psychology to examine the predictions of the theory. The results demonstrate that the expectations of listeners elicited in a range of melodic contexts may be accounted for in terms of the combined influence of a sensitivity to certain dimensions of musical events and simple, domain general learning mechanisms which are given extensive exposure to music in a given genre.

These results are significant for a number of reasons: first, they suggest an underlying cognitive account of descriptive Gestalt-based theories of expectancy in melody (*e.g.,* Narmour, 1990); second, they suggest that other cognitive accounts of music perception based on expert music-theoretic knowledge (*e.g.,* Lerdahl & Jackendoff, 1983) may significantly overestimate the perceptual and cognitive capacities of listeners. Third, they offer the possibility of addressing the bias inherently associated with such theories. Fourth, they offer the possibility of a more parsimonious model of the influences of acquired cultural influences on music perception (Cross, 1998b; Eerola, 2004b).

In Chapter 9, computational constraints on composition were examined by applying a number of multiple viewpoint systems to the task of generating successful melodies in a specified style. In spite of efforts made to improve on the modelling and sampling strategies adopted in previous research, the results demonstrated that these simple grammars are largely incapable of meeting the intrinsic demands of the task. Although the result was negative, it nonetheless remains a contribution to our understanding of cognitive processes in composition. In particular, the result is significant in the light of arguments made in previous research that similar grammars underlie the perception and composition of music (Baroni, 1999; Lerdahl, 1988a). In contention with such arguments, although the finite context grammars developed in the present research accounted rather well for a range of empirically observed phenomena in the perception of melody, they proved largely incapable of composing a stylistically successful melody. Although this may have been due in part to the MCMC sampling method used, it is noteworthy that this method represents an improvement (in terms of obtaining an unbiased sample from the target distribution) over the sequential random sampling method used in previous research. In addition, the methodology developed to evaluate the computational systems constitutes a significant contribution to future research in the cognitive modelling of composition.

Finally, the goal of applied AI is to use existing AI techniques to develop applications for specific purposes in industry. While this has not been a direct concern in the present research, the techniques developed could be put to practical use in a variety of contexts. For example, the contributions made in the present research to the statistical modelling of music and understanding of cognitive processes in music perception and composition could be put to practical use in systems for computer-assisted composition (Ames, 1989; Assayag *et al.*, 1999; Hall & Smith, 1996), machine improvisation with human performers (Lartillot *et al.*, 2001; Rowe, 1992) and music information retrieval (Pickens *et al.*, 2003).

To this extent, the present research represents an indirect contribution to such fields of applied AI.

10.3 Limitations and Future Directions

As discussed in §1.4, a number of general limitations were placed on the scope of the present research. Perhaps the most notable of these limitations is the decision to focus on monophonic music. An additional limitation arises from the fact that the results reported in this dissertation (with the partial exception of those in Chapter 6) have been obtained using a restricted set of corpora of Western folk and hymn melodies. In the case of the results reported in Chapters 6 and 7, the close alignment with results in data compression, statistical language modelling and machine learning research does suggest that the improvements afforded are robust across domains. Nonetheless, our confidence in the generality of the results would be increased if they could be experimentally replicated in a broader context of musical styles and with homophonic and polyphonic music.[1] In this regard, however, the present research does provide a basis for the formulation of specific hypotheses which may be refuted or corroborated by further experimental research.

Future research should also address some limitations in the methodology adopted in the development of the statistical models in Chapters 6 and 7. First, during the development of the statistical models used in the present research, a methodological strategy was employed whereby the best performing techniques in one experiment were adopted without further consideration in examining the performance of other dimensions of the model. As a consequence of this strategy, the resulting models will reflect local optima in the parameter space but are not guaranteed to be globally optimal.

Other limitations of the present research concern the representation scheme and features used. The strategy taken has been to use previous research in music perception and the computational modelling of music to construct by hand viewpoints which are expected to be involved in a given modelling task. As a consequence of this approach, none of the conclusions reached in the present research can be guaranteed to hold for attributes outside of the finite set used. This limitation is perhaps most notable in the case of the experiments reported in Chapters 6 and 7 (Experiment 1) where a single attribute and a single multiple viewpoint system were used respectively.

[1] Issues involved in the representation of such music for training statistical models are discussed by, for example, Assayag *et al.* (1999), Conklin (2002), Pickens *et al.* (2003) and Ponsford *et al.* (1999).

Even in later chapters, when the multiple viewpoint systems used were derived through feature selection on the basis of objective criteria, the features were drawn from a finite, hand constructed set. In addition, the features were developed for the primary purpose of modelling pitch structure. This decision may be justified by noting that pitch is typically the most complex dimension of the melodic music considered in this dissertation. However, future research should examine a wider range of features. For example, comparable analysis of a musical genre exhibiting greater degrees of rhythmic and metric complexity than the corpora used here would require the development of a rich set of viewpoints for modelling rhythmic structure. In this regard, it is worth noting that there has been relatively little psychological research conducted on temporal expectancies (*cf.* Jones & Boltz, 1989; Palmer & Krumhansl, 1990) and how these interact with melodic and tonal-harmonic factors. In such an endeavour, the multiple viewpoint framework and methodologies for evaluation employed in the present research provide a platform for the construction of hypotheses about the features represented in the cognitive processing of melody as well as the empirical examination of these hypotheses.

Further issues arise from the assumed components of the basic event space which includes attributes based on such features of the score as time signature, key signature and fermatas. Especially in the context of modelling music perception, the integration of cognitive theories of segmentation (*e.g.,* Ferrand *et al.*, 2002), tonality induction (Vos, 2000) and metre induction (*e.g.,* Eck, 2002; Toiviainen & Eerola, 2004) into the framework presented here remains a topic for future research. In addition, the construction by hand of derived features evades the important issue of how such representations may be acquired in human cognition which deserves attention in future research.

In computational terms, it would be possible to automate the construction of derived and product types using the methods developed by Lewin (1987) for constructing quotient and product GISs (see §5.2.1). In the case of derived types, a search space could be defined through the recursive application of a finite set of basic operations which partition the domain associated with a type into equivalence classes. This space could then be searched in order to optimise some objective criterion such as prediction performance. As an example, consider the types `cpitch`, `cpint`, `cpcint` and `contour` which progressively partition the pitch space into more abstract equivalence classes. Intermediate levels of abstraction (not considered in the present research) could be developed to model, for example, the hypothesised equivalence of pitches which differ by an integral number of major thirds (Conklin, 1990; Shepard, 1982). The goal

would then be to find an optimal level of abstraction in the pitch representation. Another example is provided by those types used in the present research which model pitch in relation to some referent tone such as the notated tonic (`cpintfref`), the first event in the piece (`cpintfip`) or the first event in the current bar (`cpintfib`) or phrase (`cpintfiph`). It would be possible to search a space of possible referents used in the construction of such types.

In the present research, a limited set of product types were defined between pairs of primitive types on the basis of research in music cognition and music informatics. Once again, it would be possible to search the space of possible product types between any number of primitive types to optimise some objective criterion. Finally, a space could be defined over a richer set of test types than has been examined in the current research which could then be searched in the construction of threaded types to optimise some objective criterion. The space of test types could be constrained by music-theoretic concerns as in the present research (e.g., metric accent or phrase structure) or using cognitive models of segmentation based on salient discontinuities in melodic structure. Of particular interest in this regard are models which predict segment boundaries at points where sequential expectations are non-specific or are violated (Ferrand *et al.*, 2002; Saffran *et al.*, 1999). This kind of model would fit neatly into the computational framework developed in this dissertation and the application of the computational methods developed herein to modelling the perception of segment boundaries remains an important topic for future research. While on the topic of threaded types, it remains to be seen whether the failure of finite context models to represent embedded structure and non-adjacent dependencies in the context of modelling composition (see Chapter 9) can be addressed either by threaded types or more complex types within the multiple viewpoints framework (Conklin, 2003).

A final representational issue arises from the assumption that cognitive processes operate over a symbolic musical surface consisting of discrete events. While this approach is justified to an extent by previous research in music perception and cognition, future research should focus on the relationship between symbolic and sub-symbolic levels of processing. In the present context, it is particularly important to examine perceptual constraints on the kinds of high-level symbolic attribute which may be recovered through processing of lower level representations of the musical surface (Gjerdingen, 1999b; Todd *et al.*, 1999; Šerman & Griffith, 2004).

Many opportunities exist for further development of the statistical modelling framework presented in Chapters 6 and 7. First, Bunton (1997) describes

an information-theoretic variable order state selection mechanism which replaces the original state selection used in PPM* (see §6.2.3.6) and which consistently improves performance in data compression experiments. It remains to be seen whether this mechanism can be fruitfully applied with music data. Alternatively, a variable order state selection policy based on principles of perceptual segmentation might prove profitable in the context of PPM* modelling (Reis, 1999). Second, it would be useful to use the empirical methodologies adopted in the present research to compare the performance of the PPM variants examined here with that of other modelling strategies (see §3.4 and §3.5). Examples of such techniques include: other smoothing techniques commonly used in statistical language modelling such as Katz backoff (Katz, 1987) and Kneser-Ney smoothing (Kneser & Ney, 1995); models based on the Ziv-Lempel dictionary compression algorithm (Ziv & Lempel, 1978) as used by Dubnov, Assayag and their colleagues (Assayag *et al.*, 1999; Dubnov *et al.*, 1998; Lartillot *et al.*, 2001); the PSTs (Ron *et al.*, 1996) used by Lartillot *et al.* (2001) and Triviño-Rodriguez & Morales-Bueno (2001); and the neural network models developed by Mozer (1994).

Recently, Begleiter *et al.* (2004) have examined the prediction performance of a number of variable order Markov models including PSTs, Ziv-Lempel variants and PPM. Using a methodology similar to that employed in Chapter 6, it was found that PPM (with escape method C) and the Context Tree Weighting (CTW) compression algorithm (Willems *et al.*, 1995) outperform the other algorithms in predicting sequential data drawn from three different domains: molecular biology, text and music.[2] While these results provide convergent evidence for the relative power of the PPM modelling strategies used in the present research, they also suggest that further examination of the CTW algorithm may prove fruitful in future research.

It would also be useful to conduct a thorough examination of the effect of the overall architecture of the multiple viewpoint framework on performance. How is performance affected, for example, if we first combine the LTM-STM predictions for each viewpoint and then combine the resulting predictions? It seems unlikely that a single combination of all distributions will improve performance but this conjecture can only be tested by empirical experimentation. In addition, it remains to be seen whether other combination schemes developed in the field of machine learning (*e.g.,* Chen *et al.*, 1997; Kittler *et al.*, 1998;

[2]It should be noted that Begleiter *et al.* (2004) use music as a source of complex real-world sequential data for evaluating the performance of general-purpose algorithms. They do not examine issues particular to the cognitive or analytical representation and processing of musical structure.

Xu *et al.*, 1992) can be profitably applied to modelling music with multiple viewpoint systems. Finally, it was suggested in §10.2 that the modelling techniques evaluated in the present research might be used profitably in practical applications such as tools for composition, performance and music information retrieval. In order to address such questions, future research should examine the validity as entropy as a measure of performance through detailed empirical studies of the relationship between entropy measures and model performance on these practical tasks. Other methodological issues to be considered would concern the effects of training set size and homogeneity on performance and reliability (Knopoff & Hutchinson, 1983).

In Chapter 8, a theory of expectancy in music was proposed and empirically evaluated in the context of observed patterns of expectation collected in previous research. Future developments should extend this approach to a more comprehensive examination of statistical learning-based systems in the context of continuous response methodologies (Eerola *et al.*, 2002; Schubert, 2001) where expectancies are elicited throughout listening to a piece of music. Within such an approach, a detailed examination of any changes in the relative importance of different features over time would provide interesting data on dynamic aspects of expectancy. A start could be made in this direction by examining the weights assigned to long and short-term models of different features over time. In addition to examining the relationship between the responses of listeners and those of the model, future work should also focus on examining relationships between these responses and objective musical features as well as potential neurophysiological correlates of expectancy (Koelsch *et al.*, 2000).

According to the proposed theory of expectancy, the musical experience of listeners will have an effect both on the observed patterns of expectation and the features which influence those patterns. In fact, the theory makes some predictions about the influences of musical exposure that should be tested experimentally. It would be predicted, for example, that a model trained on the music of one culture would predict the expectations of people of that culture better than a model trained on the music of another culture and *vice versa* (see also Castellano *et al.*, 1984). Although, further research is required to examine these predictions, it is expected that the learning based theory will be able to account more parsimoniously than existing rule based theories (Narmour, 1990) for both similarities and differences in observed patterns of expectation between two cultures on the basis of similarities and differences in the music of those cultures.

One of the advantages of the proposed theory of expectancy is that it can

be applied to the modelling of developmental trajectories in the acquisition of mature patterns of expectancy (Schellenberg *et al.*, 2002). Research along these lines would be capable of examining the basic representational capacities of infants, how these develop, how new representations are acquired through increasing exposure and how this process relates to general cognitive development and the acquisition of other skills such as language.

While the experimental results support the theory of expectancy at the computational level, further investigation is required to analyse the model at finer levels of description, to generate hypotheses at these levels and to subject the hypotheses to experimental evaluation. Such hypotheses might concern, for example, constraints placed on model order by human working memory limitations, the interaction of the short- and long-term models and the effects of intra- and extra-opus experience as well as more detailed proposals concerning the manner in which regularities in different melodic features contribute to melodic expectancy. Finally, experiments with more complex polyphonic contexts may find the systems developed in the present research to be underspecified, while continued research with melodic contexts may find the present systems to be overspecified in some respects.

In Chapter 9, the finite context systems, which proved highly successful in modelling aspects of perceptual expectancy, failed to meet the intrinsic demands of the task of composing a successful chorale melody. Nonetheless the experimental approach allowed an examination of some limitations of the systems examined and their subsequent improvement. While the preliminary results are promising (see Appendix G), the improved models should be fully examined using the methodology developed in the present research to identify in which areas of the task they succeed, if any, and in which they fail and subsequently elaborated on the basis of these findings. While the examination of more powerful grammars was beyond the scope of the present research, these should form a key part of future efforts to examine computational constraints on composition. In particular, the use of short-term models was identified as an insufficient means of maintaining coherence in intra-opus structure. Future work might address this limitation though the use of symbolic constraints derived from music theory or music perception (Hall & Smith, 1996; Hild *et al.*, 1992; Povel, 2004) or through more sophisticated representational structures and pattern-based stochastic sampling (Conklin, 2003). The methodological framework developed in the present research represents the beginnings of a framework for the detailed, empirical examination and comparison of theories of cognitive processing in composition, the refutation and corroboration of such

theories at varying levels of description and their subsequent modification and elaboration.

APPENDIX A

NOTATIONAL CONVENTIONS

Sets

$\|S\|$	the cardinality of set S
2^S	the power set of set S
$S \times S'$	the Cartesian product of sets S and S'

Frequently Encountered Sets

$\mathbb{R}$	real numbers
$\mathbb{Q}$	rational numbers
$\mathbb{Q}^+$	positive rational numbers
$\mathbb{Q}^*$	non-negative rational numbers
$\mathbb{Z}$	integers
$\mathbb{Z}^+$	positive integers
$\mathbb{Z}^*$	non-negative integers

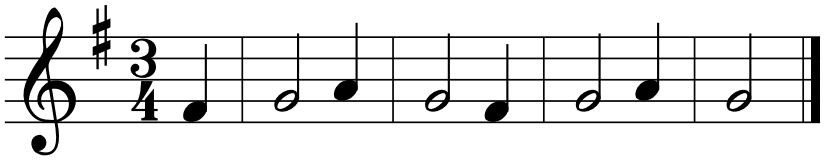
Symbols and Sequences

$\perp$	the null symbol
ε	the empty sequence
$\mathcal{A}^+$	the set of all non-empty sequences composed from elements of the alphabet $\mathcal{A}$ (the *positive closure* of $\mathcal{A}$)
$\mathcal{A}^*$	$\mathcal{A}^+ \bigcup \{\varepsilon\}$ (the *Kleene closure* of $\mathcal{A}$)
$a_i^j \in \mathcal{A}^*$	a sequence of symbols drawn from alphabet $\mathcal{A}$ indexed from i to j, $j \geq i \in \mathbb{Z}^+$
$a_i,\ i \in \mathbb{Z}^+$	the symbol at index i of sequence a_k^j
$a_i^j a_k^l$	the concatenation of two sequences

APPENDIX B

AN EXAMPLE KERN FILE

The file `jugos104.krn` in the Essen Folk Song Collection (Schaffrath, 1995) contains the Yugoslavian folksong *Das neue Grab Es hot shi der maren in an naien Grube*. This melody appears in Dataset 4 and was chosen for illustrative purposes since it is the shortest melody in the datasets used in the present research (see Chapter 4). It is, in fact, the shortest melody in the entire EFSC. Figure B.1 shows this folksong in standard music notation and as viewpoint sequences for each of the attribute types making up the basic event space used in the present research (see §5.3). The original `**kern` encoding is shown below.



`onset`	48	72	120	144	192	216	264	288
`deltast`	0	0	0	0	0	0	0	0
`dur`	24	48	24	48	24	48	24	48
`barlength`	72	72	72	72	72	72	72	72
`pulses`	3	3	3	3	3	3	3	3
`cpitch`	66	67	69	67	66	67	69	67
`keysig`	1	1	1	1	1	1	1	1
`mode`	0	0	0	0	0	0	0	0
`phrase`	1	0	0	0	0	0	0	-1

Figure B.1: An example melody from the EFSC.

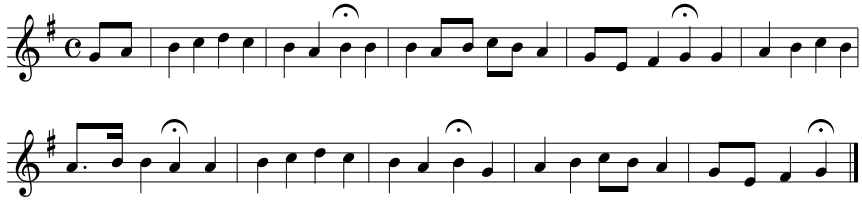
```
!!!OTL: Das neue Grab Es hot shi der maren in an naien Grube:
!!!ARE: Europa, Suedosteuropa, Jugoslavija, Gottschee, Ober Wetzenbach
!!!YOR: 5, S. 226
!!    1909 aufgezeichnet
!!!SCT: Q0110A
!!!YEM: Copyright 1995, estate of Helmut Schaffrath.
**kern
*ICvox
*Ivox
*M3/4
*k[f#]
*G:
{4f#
=1
2g
4a
=2
2g
4f#
=3
2g
4a
=4
2g}
==
!!!AGN: Ballade, Tod, Geburt
!!!ONB: ESAC (Essen Associative Code) Database: BALLADE
!!!AMT: simple triple
!!!AIN: vox
!!!EED: Helmut Schaffrath
!!!EEV: 1.0
*-
```

APPENDIX C

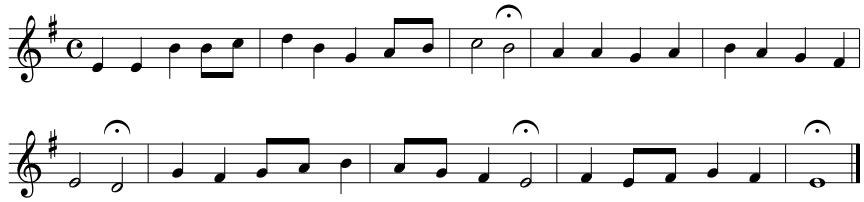
SEVEN ORIGINAL CHORALE MELODIES

This appendix contains the seven original chorale melodies used as a basis for the experiments presented in Chapter 9. The numbering system is that of Riemenschneider (1941) and BWV numbers are given in brackets after the title of each chorale. Note that repeated sections, which occur in chorales 249, 365 and 44, are not expanded as discussed in Chapter 5.

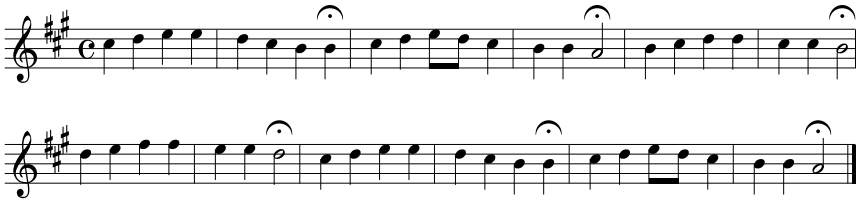
249: *Allein Gott in der Höh' sei Ehr* (BWV 260)



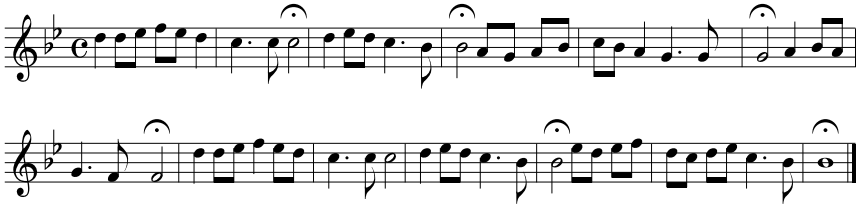
238: *Es wird schier der letzte Tag herkommen* (BWV 310)



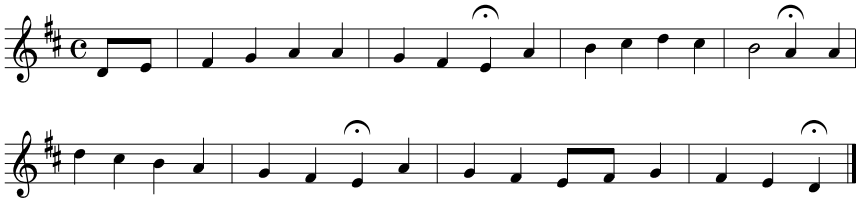
365: *Jesu, meiner Seelen Wonne* (BWV 359)



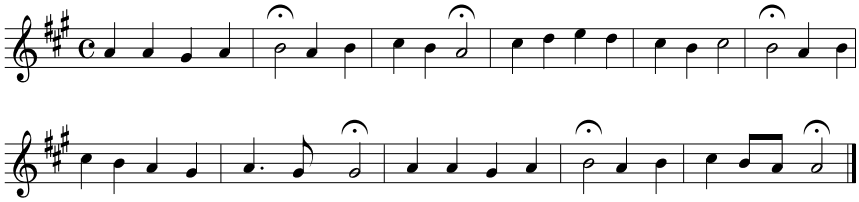
264: *Jesu, meines Herzens Freud'* (BWV 361)



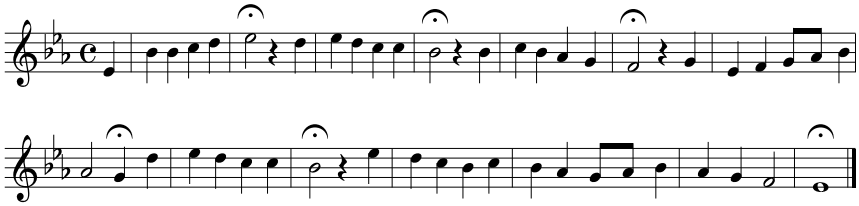
44: *Mach's mit mir, Gott, nach deiner Güt* (BWV 377)



141: *Seelenbräutigam, Jesu, Gottes Lamm* (BWV 409)



147: *Wenn ich in Angst und Not* (BWV 427)

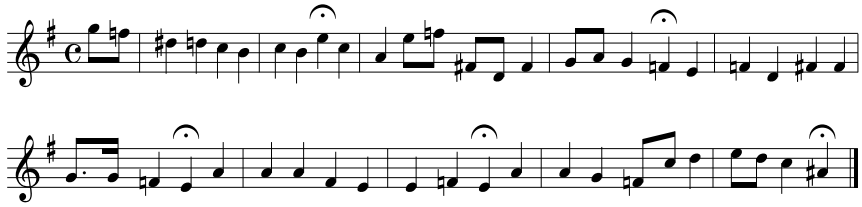


APPENDIX D

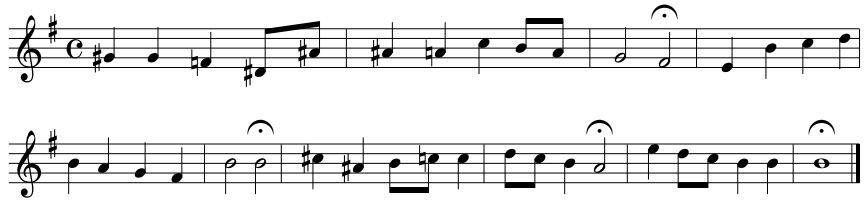
MELODIES GENERATED BY SYSTEM A

This appendix contains the melodies generated by System A as discussed in Chapter 9. Each melody is numbered and titled according to the original melody on which it is based and from which it derives its time signature, key signature, rhythmic structure and phrase structure. Appendix C contains the seven original chorale melodies used as a basis for the generation of these melodies.

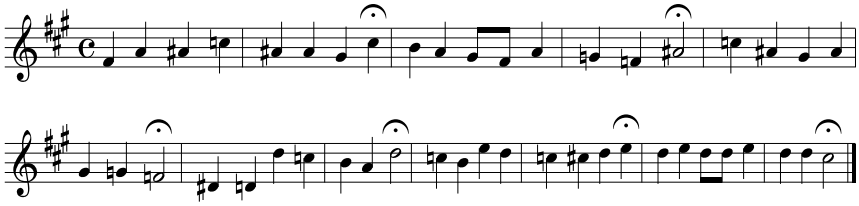
A249: *Allein Gott in der Höh' sei Ehr*



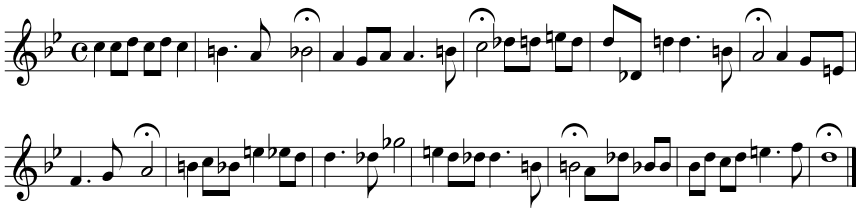
A238: *Es wird schier der letzte Tag herkommen*



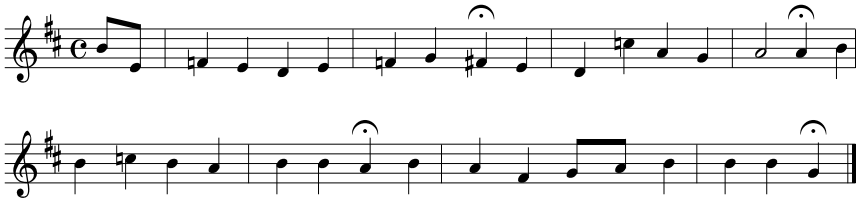
A365: *Jesu, meiner Seelen Wonne*



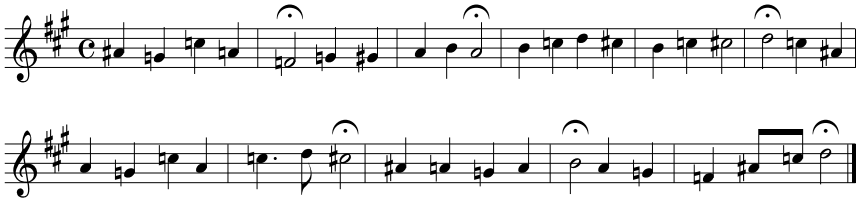
A264: *Jesu, meines Herzens Freud'*



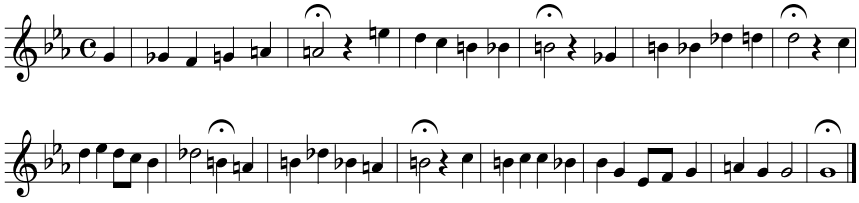
A44: *Mach's mit mir, Gott, nach deiner Güt*



A141: *Seelenbräutigam, Jesu, Gottes Lamm*



A147: *Wenn ich in Angst und Not*

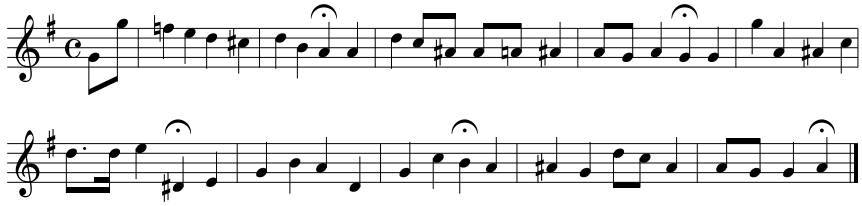


APPENDIX E

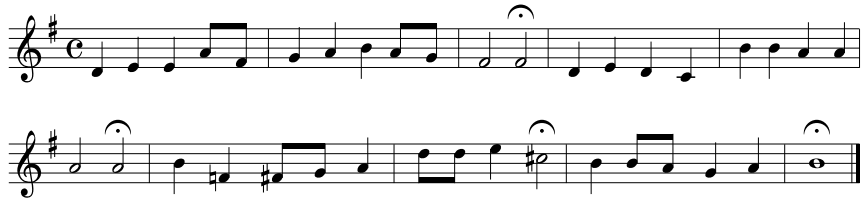
MELODIES GENERATED BY SYSTEM B

This appendix contains the melodies generated by System B as discussed in Chapter 9. Each melody is numbered and titled according to the original melody on which it is based and from which it derives its time signature, key signature, rhythmic structure and phrase structure. Appendix C contains the seven original chorale melodies used as a basis for the generation of these melodies.

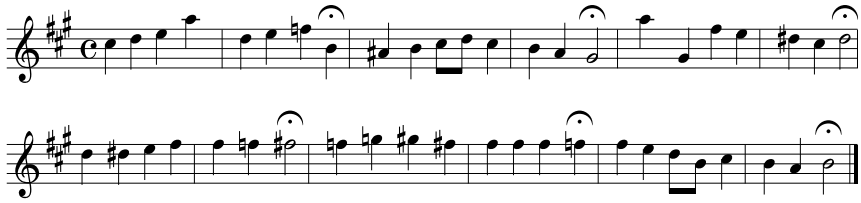
B249: *Allein Gott in der Höh' sei Ehr*



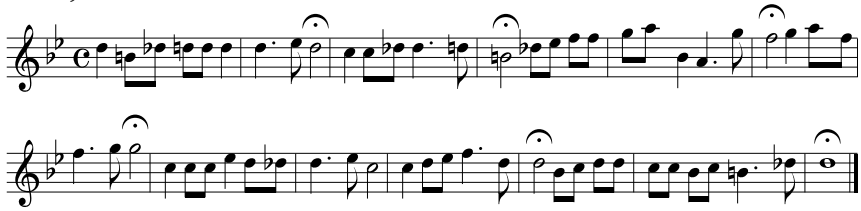
B238: *Es wird schier der letzte Tag herkommen*



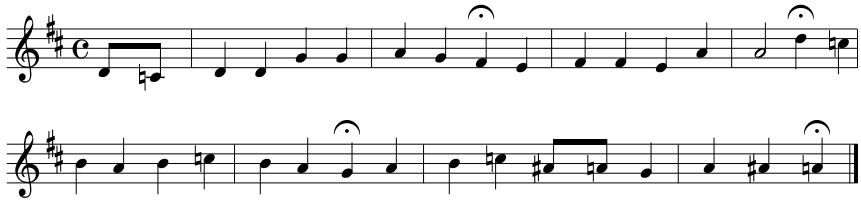
B365: *Jesu, meiner Seelen Wonne*



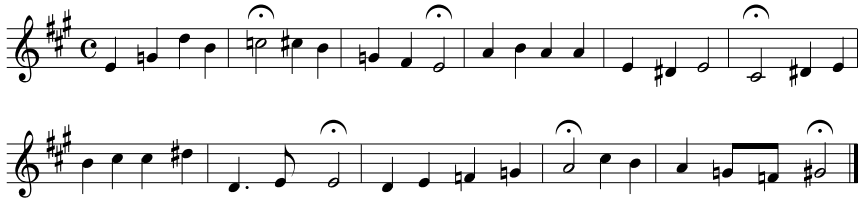
B264: *Jesu, meines Herzens Freud'*



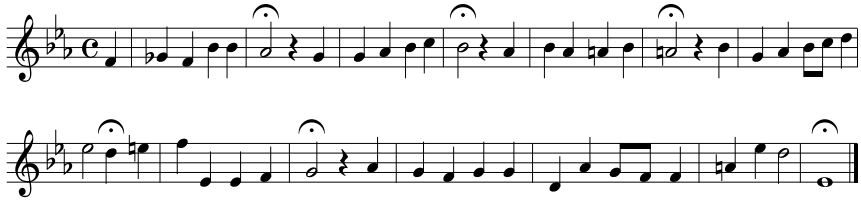
B44: *Mach's mit mir, Gott, nach deiner Güt*



B141: *Seelenbräutigam, Jesu, Gottes Lamm*



B147: *Wenn ich in Angst und Not*

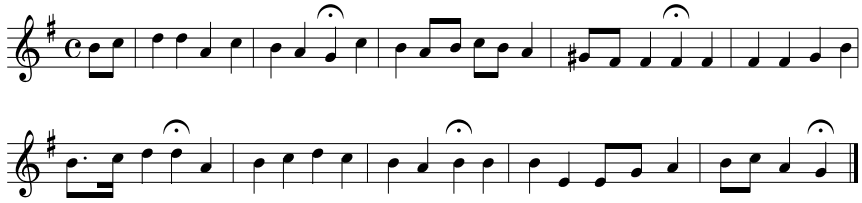


APPENDIX F

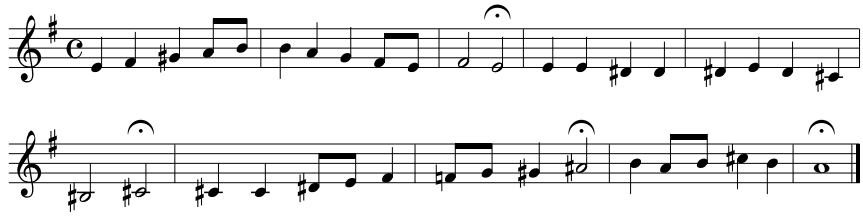
MELODIES GENERATED BY SYSTEM C

This appendix contains the melodies generated by System C as discussed in Chapter 9. Each melody is numbered and titled according to the original melody on which it is based and from which it derives its time signature, key signature, rhythmic structure and phrase structure. Appendix C contains the seven original chorale melodies used as a basis for the generation of these melodies.

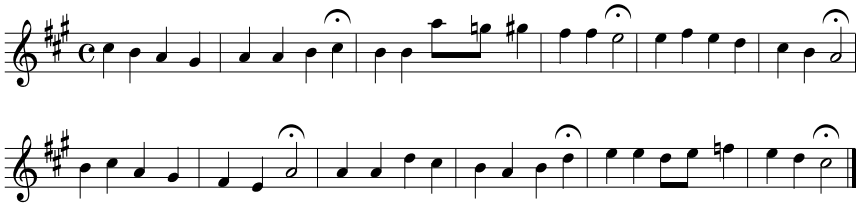
C249: *Allein Gott in der Höh' sei Ehr*



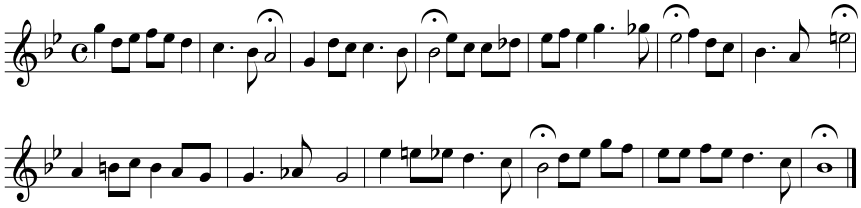
C238: *Es wird schier der letzte Tag herkommen*



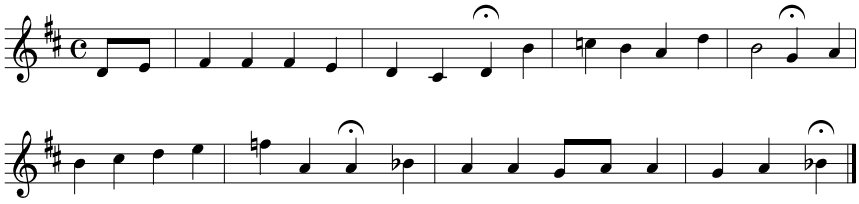
C365: *Jesu, meiner Seelen Wonne*



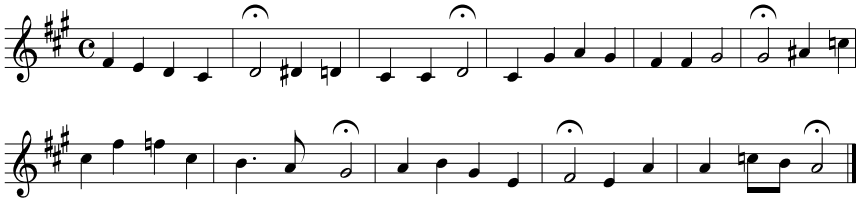
C264: *Jesu, meines Herzens Freud'*



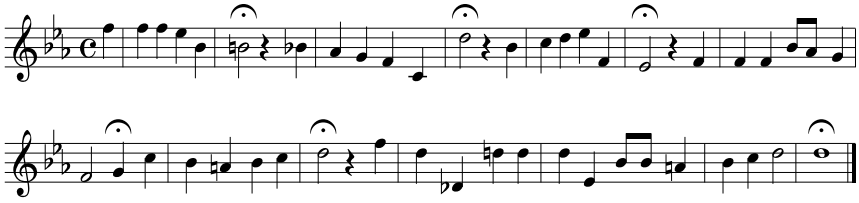
C44: *Mach's mit mir, Gott, nach deiner Güt*



C141: *Seelenbräutigam, Jesu, Gottes Lamm*



C147: *Wenn ich in Angst und Not*



APPENDIX G

A MELODY GENERATED BY SYSTEM D

This appendix contains the results of a preliminary investigation into the capacity of System D to generate stylistically successful chorale melodies. As discussed in §9.5.5, System D was derived through feature selection to reduce entropy over Dataset 2 (see Table 4.1) using a feature set augmented with additional viewpoints in order to address the failure of System C to represent some salient stylistic constraints of the corpus. System D comprises the viewpoints shown in Table 9.8 and exhibits significantly lower entropy than System C in predicted unseen melodies in Dataset 2. System D was used to generate a several melodies following the procedure described in §9.4.3 with the seven chorale melodies shown in Appendix C used as base melodies.

Figure G.1 shows the most successful melody generated by System D using Chorale 365 as its base melody. In terms of tonal and melodic structure, it is much more coherent than the melodies generated by System C. The multiple regression model developed in §9.5.4 to account for the judges' ratings of stylistic success predict that this melody would receive a rating of 6.4 on

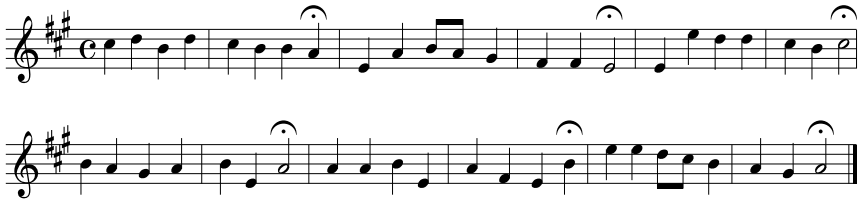


Figure G.1: Chorale D365 generated by System D.

a seven-point scale of success as a chorale melody. While these preliminary results are encouraging, the remaining melodies generated were less successful and System D must be fully analysed using the methodology developed in Chapter 9 in order to examine its ability to *consistently* compose original and stylistically successful melodies.

BIBLIOGRAPHY

Aarden, B. (2003). *Dynamic Melodic Expectancy*. Doctoral dissertation, Ohio State University, Columbus, OH.

Aha, D. W. & Bankert, R. L. (1996). A comparative evaluation of sequential feature selection algorithms. In D. Fisher & H. J. Lenz (Eds.), *Learning from Data: AI and Statistics V* (pp. 199–206). New York: Springer.

Allan, M. (2002). *Harmonising Chorales in the Style of J. S. Bach*. Master's dissertation, School of Informatics, University of Edinburgh, UK.

Amabile, T. M. (1996). *Creativity in Context*. Boulder, Colorado: Westview Press.

Ames, C. (1987). Automated composition in retrospect: 1956–1986. *Leonardo*, *20*(2), 169–185.

Ames, C. (1989). The Markov process as a compositional model: A survey and tutorial. *Leonardo*, *22*(2), 175–187.

Ames, C. (1992). Quantifying musical merit. *Interface*, *21*(1), 53–93.

Ames, C. & Domino, M. (1992). Cybernetic Composer: An overview. In M. Balaban, K. Ebcioğlu, & O. Laske (Eds.), *Understanding Music with AI: Perspectives on Music Cognition* (pp. 186–205). Cambridge, MA: MIT Press.

Anderson, J. R. (2000). *Learning and Memory: An Integrated Approach* (Second ed.). New York: John Wiley & Sons, Inc.

Assayag, G., Dubnov, S., & Delerue, O. (1999). Guessing the composer's mind: Applying universal prediction to musical style. In *Proceedings of the 1999 International Computer Music Conference* (pp. 496–499). San Francisco: ICMA.

Assayag, G., Rueda, C., Laurson, M., Agon, C., & Delerue, O. (1999). Computer assisted composition at IRCAM: From PatchWork to OpenMusic. *Computer Music Journal*, *23*(3), 59–72.

Auh, M. (2000). Assessing creativity in composing music: Product–process–person–environment approaches. In *Proceedings of the 2000 National Conference of the Australian Association for Research in Education*. Sydney, Australia: UTS.

Auh, M. & Johnston, R. (2000). A pilot study of comparing creativity in composing and story–telling by children aged 3 to 5. In *Proceedings of the 2000 UTS Research Symposium* (pp. 21–27). Sydney, Australia: UTS.

Auh, M. & Walker, R. (1999). Compositional strategies and musical creativity when composing with staff notations versus graphic notations among Korean students. *Bulletin of the Council for Research in Music Education*, *141*, 2–9.

Balzano, G. J. (1982). The pitch set as a level of description for studying musical pitch perception. In M. Clynes (Ed.), *Music, Mind and Brain* (pp. 321–351). New York: Plenum.

Balzano, G. J. (1986a). Music perception as detection of pitch–time constraints. In V. McCabe & G. J. Balzano (Eds.), *Event Cognition: An Ecological Perspective* (pp. 217–233). Hillsdale, NJ: Erlbaum.

Balzano, G. J. (1986b). What are musical pitch and timbre? *Music Perception*, *3*(3), 297–314.

Baroni, M. (1999). Musical grammar and the cognitive processes of composition. *Musicæ Scientiæ*, *3*(1), 3–19.

Baroni, M., Dalmonte, R., & Jacoboni, C. (1992). Theory and analysis of European melody. In A. Marsden & A. Pople (Eds.), *Computer Representations and Models in Music* (pp. 187–206). London: Academic Press.

Begleiter, R., El-Yaniv, R., & Yona, G. (2004). On prediction using variable order Markov models. *Journal of Artificial Intelligence Research*, *22*, 385–421.

Bell, T. C., Cleary, J. G., & Witten, I. H. (1990). *Text Compression*. Englewood Cliffs, NJ: Prentice Hall.

Bergeson, T. R. (1999). Melodic expectancy in infancy. *Journal of the Acoustical Society of America*, *106*(4), 2285.

Berlyne, D. E. (1974). The new experimental aesthetics. In D. E. Berlyne (Ed.), *Studies in the New Experimental Aesthetics: Steps Towards an Objective Psychology of Aesthetic Appreciation* (pp. 1–25). Washington: Hemisphere Publishing Co.

Bharucha, J. J. (1984). Anchoring effects in music: The resolution of dissonance. *Cognitive Psychology*, *16*, 485–518.

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 P. Todd & G. Loy (Eds.), *Music and Connectionism* (pp. 84–99). Cambridge, MA: MIT Press.

Bharucha, J. J. (1993). Tonality and expectation. In R. Aiello (Ed.), *Musical Perceptions* (pp. 213–239). Oxford: Oxford University Press.

Bharucha, J. J. & Stoeckig, K. (1986). Reaction time and musical expectancy: Priming of chords. *Journal of Experimental Psychology: Human Perception and Performance*, *12*(4), 403–410.

Bharucha, J. J. & Stoeckig, K. (1987). Priming of chords: Spreading activation or overlapping frequency spectra? *Perception and Psychophysics*, *41*(6), 519–524.

Bharucha, J. J. & Todd, P. (1989). Modelling the perception of tonal structure with neural nets. *Computer Music Journal*, *13*(4), 44–53.

Blum, A. & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, *97*(1–2), 245–271.

Boden, M. A. (1990). *The Creative Mind: Myths and Mechanisms*. London: Weidenfield and Nicholson.

Boltz, M. G. (1989a). Rhythm and "good endings": Effects of temporal structure on tonality judgements. *Perception and Psychophysics*, *46*(1), 9–17.

Boltz, M. G. (1989b). Time judgements of musical endings: Effects of expectancies on the 'filled interval effect'. *Perception and Psychophysics*, *46*(5), 409–418.

Boltz, M. G. (1993). The generation of temporal and melodic expectancies during musical listening. *Perception and Psychophysics*, *53*(6), 585–600.

Boltz, M. G. & Jones, M. R. (1986). Does rule recursion make melodies easier to reproduce? If not, what does? *Cognitive Psychology*, *18*(4), 389–431.

Brinkman, D. J. (1999). Problem finding, creativity style and the musical compositions of high school students. *Journal of Creative Behaviour*, *33*(1), 62–68.

Brochard, R., Dufour, A., Drake, C., & Scheiber, C. (2000). Functional brain imaging of rhythm perception. In C. Woods, G. Luck, R. Brochard, F. Seddon, & J. A. Sloboda (Eds.), *Proceedings of the Sixth International Conference of Music Perception and Cognition*. Keele, UK: University of Keele.

Brooks Jr., F. P., Hopkins, A. L., Neumann, P. G., & Wright, W. V. (1957). An experiment in musical composition. *IRE Transactions on Electronic Computers*, *EC-6*(1), 175–182.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Lai, J. C., & Mercer, R. L. (1992). An estimate of an upper bound on the entropy of English. *Computational Linguistics*, *18*(1), 32–40.

Brown, R. T. (1989). Creativity: What are we to measure? In J. Glover, R. Ronning, & C. Reynolds (Eds.), *Handbook of Creativity* (pp. 3–32). New York: Plenum Press.

Bundy, A. (1990). What kind of field is AI? In D. Partridge & Y. Wilks (Eds.), *The Foundations of Artificial Intelligence* (pp. 215–222). Cambridge, UK: Cambridge University Press.

Bunton, S. (1996). *On-Line Stochastic Processes in Data Compression*. Doctoral dissertation, University of Washington, Seattle, WA.

Bunton, S. (1997). Semantically motivated improvements for PPM variants. *The Computer Journal*, *40*(2/3), 76–93.

Cambouropoulos, E. (1996). A general pitch interval representation: Theory and applications. *Journal of New Music Research*, *25*(3), 231–251.

Cambouropoulos, E. (1998). *Towards a General Computational Theory of Musical Structure*. Doctoral dissertation, The University of Edinburgh, Faculty of Music and Department of Artificial Intelligence.

Cambouropoulos, E., Crawford, T., & Iliopoulos, C. S. (1999). Pattern processing in melodic sequences: Challenges, caveats and prospects. In *Proceedings of the AISB'99 Symposium on Musical Creativity* (pp. 42–47). Brighton, UK: SSAISB.

Camilleri, L. (1992). Computational theories of music. In A. Marsden & A. Pople (Eds.), *Computer Representations and Models in Music* (pp. 171–185). London: Academic Press.

Carlsen, J. C. (1981). Some factors which influence melodic expectancy. *Psychomusicology*, *1*(1), 12–29.

Castellano, M. A., Bharucha, J. J., & Krumhansl, C. L. (1984). Tonal hierarchies in the music of North India. *Journal of Experimental Psychology: General*, *113*(3), 394–412.

Chalmers, D. J. (1994). On implementing a computation. *Minds and Machines*, *4*, 391–402.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organisation. *Psychological Review*, *103*(3), 566–581.

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *The Quarterly Journal of Experimental Psychology*, *52A*(2), 273–302.

Chater, N. & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22.

Chen, K., Wang, L., & Chi, H. (1997). Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification. *International Journal of Pattern Recognition and Artificial Intelligence*, *11*(3), 417–415.

Chen, S. F. & Goodman, J. (1999). An empirical study of smoothing techniques for language modelling. *Computer Speech and Language*, *13*(4), 359–394.

Clark, D. M. & Wells, A. (1997). Cognitive therapy for anxiety disorders. In L. J. Dickstein, M. B. Riba, & J. M. Oldham (Eds.), *Review of Psychiatry*, volume 16. American Psychiatric Press.

Clarke, E. F. (1985). Some aspects of rhythm and expression in performances of Erik Satie's "Gnossienne No. 5". *Music Perception*, *2*(3), 299–328.

Cleary, J. G. & Teahan, W. J. (1995). Some experiments on the zero-frequency problem. In J. A. Storer & M. Cohn (Eds.), *Proceedings of the IEEE Data Compression Conference*. Washington, DC: IEEE Computer Society Press.

Cleary, J. G. & Teahan, W. J. (1997). Unbounded length contexts for PPM. *The Computer Journal*, *40*(2/3), 67–75.

Cleary, J. G. & Witten, I. H. (1984). Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, *32*(4), 396–402.

Cohen, A. J. (2000). Development of tonality induction: Plasticity, exposure and training. *Music Perception*, *17*(4), 437–459.

Cohen, A. J., Cuddy, L. L., & Mewhort, D. J. K. (1977). Recognition of transposed tone sequences. *Journal of the Acoustical Society of America*, *61*, 87–88.

Colley, A., Banton, L., Down, J., & Pither, A. (1992). An expert–novice comparison in musical composition. *Psychology of Music*, *20*(2), 124–137.

Conklin, D. (1990). *Prediction and Entropy of Music*. Master's dissertation, Department of Computer Science, University of Calgary, Canada.

Conklin, D. (2002). Representation and discovery of vertical patterns in music. In C. Anagnostopoulou, M. Ferrand, & A. Smaill (Eds.), *Proceedings of the Second International Conference of Music and Artificial Intelligence*, volume 2445 of *Lecture Notes in Computer Science* (pp. 32–42). Berlin: Springer.

Conklin, D. (2003). Music generation from statistical models. In *Proceedings of the AISB 2003 Symposium on Artificial Intelligence and Creativity in the Arts and Sciences* (pp. 30–35). Brighton, UK: SSAISB.

Conklin, D. & Anagnostopoulou, C. (2001). Representation and discovery of multiple viewpoint patterns. In *Proceedings of the 2001 International Computer Music Conference*. San Francisco: ICMA.

Conklin, D. & Cleary, J. G. (1988). Modelling and generating music using multiple viewpoints. In *Proceedings of the First Workshop on AI and Music* (pp. 125–137). Menlo Park, CA: AAAI Press.

Conklin, D. & Witten, I. H. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research*, *24*(1), 51–73.

Conley, J. K. (1981). Physical correlates of the judged complexity of music by subjects differing in musical background. *British Journal of Psychology*, *72*(4), 451–464.

Cook, N. (1987). The perception of large-scale tonal closure. *Music Perception*, *5*(2), 197–206.

Cook, N. (1994). Perception: A perspective from music theory. In R. Aiello & J. Sloboda (Eds.), *Musical Perceptions* (pp. 64–95). Oxford: Oxford University Press.

Cope, D. (1991). *Computers and Musical Style*. Oxford: Oxford University Press.

Cope, D. (1992a). Computer modelling of musical intelligence in EMI. *Computer Music Journal*, *16*(2), 69–83.

Cope, D. (1992b). On algorithmic representation of musical style. In M. Balaban, K. Ebcioğlu, & O. Laske (Eds.), *Understanding Music with AI: Perspectives on Music Cognition* (pp. 354–363). Cambridge, MA: MIT Press.

Cope, D. (2001). *Virtual Music*. Cambridge, MA: MIT Press.

Cover, T. M. & King, R. C. (1978). A convergent gambling estimate of the entropy of English. *IEEE Transactions on Information Theory*, *24*(4), 413–421.

Creighton, H. (1966). *Songs and Ballads from Nova Scotia*. New York: Dover.

Cross, I. (1995). Review of *The analysis and cognition of melodic complexity: The implication-realization model*, Narmour (1992). *Music Perception*, *12*(4), 486–509.

Cross, I. (1998a). Music analysis and music perception. *Music Analysis*, *17*(1), 3–20.

Cross, I. (1998b). Music and science: Three views. *Revue Belge de Musicologie*, *52*, 207–214.

Cuddy, L. L. & Lunny, C. A. (1995). Expectancies generated by melodic intervals: Perceptual judgements of continuity. *Perception and Psychophysics*, *57*(4), 451–462.

Cutting, J. E., Bruno, N., Brady, N. P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgements of perceived depth. *Journal of Experimental Psychology: General*, *121*(3), 364–381.

Davidson, L. & Welsh, P. (1988). From collections to structure: The developmental path of tonal thinking. In J. A. Sloboda (Ed.), *Generative Processes in Music: The Psychology of Performance, Improvisation and Composition* (pp. 260–285). Oxford: Clarendon Press.

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

Desain, P., Honing, H., van Thienen, H., & Windsor, L. (1998). Computational modelling of music cognition: Problem or solution. *Music Perception*, *16*(1), 151–166.

Deutsch, D. (1982). The processing of pitch combinations. In D. Deutsch (Ed.), *Psychology of Music* (pp. 271–316). New York: Academic Press.

Deutsch, D. & Feroe, J. (1981). The internal representation of pitch sequences in tonal music. *Psychological Review*, *88*(6), 503–522.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1924.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *First International Workshop on Multiple Classifier Systems* (pp. 1–15). New York: Springer.

Dietterich, T. G. & Michalski, R. S. (1986). Learning to predict sequences. In R. S. Michalski, J. Carbonell, & T. M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach*, volume II (pp. 63–106). San Mateo, CA: Morgan Kauffman.

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

Dowling, W. J. (1994). Melodic contour in hearing and remembering melodies. In R. Aiello & J. Sloboda (Eds.), *Musical Perceptions* (pp. 173–190). Oxford: Oxford University Press.

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

Dubnov, S., Assayag, G., & El-Yaniv, R. (1998). Universal classification applied to musical sequences. In *Proceedings of the 1998 International Computer Music Conference* (pp. 332–340). San Francisco: ICMA.

Ebcioğlu, K. (1988). An expert system for harmonising four–part chorales. *Computer Music Journal*, *12*(3), 43–51.

Eck, D. (2002). Finding downbeats with a relaxation oscillator. *Psychological Research*, *66*(1), 18–25.

Eerola, T. (2004a). Data-driven influences on melodic expectancy: Continuations in North Sami Yoiks rated by South African traditional healers. In S. D. Lipscomb, R. Ashley, R. O. Gjerdingen, & P. Webster (Eds.), *Proceedings of the Eighth International Conference of Music Perception and Cognition* (pp. 83–87). Adelaide, Australia: Causal Productions.

Eerola, T. (2004b). *The Dynamics of Musical Expectancy: Cross-cultural and Statistical Approaches to Melodic Expectations*. Doctoral dissertation, Faculty of Humanities, University of Jyväskylä, Finland. Jyväskylä Studies in Humanities, 9.

Eerola, T. & North, A. C. (2000). Expectancy-based model of melodic complexity. In C. Woods, G. Luck, R. Brochard, F. Seddon, & J. A. Sloboda (Eds.), *Proceedings of the Sixth International Conference on Music Perception and Cognition*. Keele, UK: Keele University.

Eerola, T., Toiviainen, P., & Krumhansl, C. L. (2002). Real-time prediction of melodies: Continuous predictability judgements and dynamic models. In C. Stevens, D. Burnham, E. Schubert, & J. Renwick (Eds.), *Proceedings of the Seventh International Conference on Music Perception and Cognition* (pp. 473–476). Adelaide, Australia: Causal Productions.

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

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Ferrand, M., Nelson, P., & Wiggins, G. (2002). A probabilistic model for melody segmentation. In *Electronic Proceedings of the Second International Conference on Music and Artificial Intelligence*. University of Edinburgh, UK: Springer.

Folkestad, G., Hargreaves, D. J., & Lindström, B. (1990). A typology of compositional styles in computer–based music making. In A. Gabrielsson (Ed.), *Proceedings of the Third Triennial ESCOM Conference* (pp. 147–152). Uppsala, Sweden: Uppsala University Press.

Genest, C. & Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, *1*(1), 114–148.

Ghahramani, Z. & Jordan, M. (1997). Factorial hidden Markov models. *Machine Learning*, *29*, 245–275.

Gjerdingen, R. (1999a). An experimental music theory? In N. Cook & M. Everist (Eds.), *Rethinking Music* (pp. 161–170). Oxford: Oxford University Press.

Gjerdingen, R. O. (1999b). Apparent motion in music? In N. Griffith & P. M. Todd (Eds.), *Musical Networks: Parallel Distributed Perception and Performance* (pp. 141–173). Cambridge, MA: MIT Press/Bradford Books.

Gould, S. J. (1985). *The Flamingo's Smile*. London: W. W. Norton.

Gusfield, D. (1997). *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge, UK: Cambridge University Press.

Hall, M. & Smith, L. (1996). A computer model of blues music and its evaluation. *Journal of the Acoustical Society of America*, *100*(2), 1163–1167.

Hall, M. A. (1995). Selection of attributes for modeling Bach chorales by a genetic algorithm. In *Proceedings of the Second New Zealand Two-Stream Conference on Artificial Neural Networks and Expert Systems*. Dunedin, New Zealand: IEEE Computer Society Press.

Hall, R. A. (1953). Elgar and the intonation of British English. *Gramophone*, *31*, 6.

Harris, M., Smaill, A., & Wiggins, G. (1991). Representing music symbolically. In A. Camurri & C. Canepa (Eds.), *IX Colloquio di Informatica Musicale*. Genova, Italy: Universita di Genova.

Hickey, M. (2000). The use of consensual assessment in the evaluation of children's music compositions. In C. Woods, G. Luck, R. Brochard, F. Seddon, & J. A. Sloboda (Eds.), *Proceedings of the Sixth International Conference on Music Perception and Cognition*. Keele, UK: Keele University.

Hickey, M. (2001). An application of Amabile's Consensual Assesment Technique for rating the creativity of children's musical compositions. *Journal of Research in Music Education*, *49*(3), 234–244.

Hild, H., Feulner, J., & Menzel, D. (1992). HARMONET: A neural net for harmonising chorales in the style of J. S. Bach. In R. P. Lippmann, J. E. Moody,

& D. S. Touretzky (Eds.), *Advances in Neural Information Processing 4* (pp. 267–274). San Francisco: Morgan Kaufmann.

Hiller, L. (1970). Music composed with computers – a historical survey. In H. B. Lincoln (Ed.), *The Computer and Music* (pp. 42–96). Cornell, USA: Cornell University Press.

Hiller, L. & Isaacson, L. (1959). *Experimental Music*. New York: McGraw–Hill.

Hinton, G. (1999). Products of experts. In *Proceedings of the Ninth International Conference on Artificial Neural Networks*, volume 1 (pp. 1–6). London, UK: IEE.

Hinton, G. (2000). *Training Products of Experts by Minimizing Contrastive Divergence*, (Technical Report No. GCNU TR 2000-004). Gatsby Computational Neuroscience Unit, University College London.

Hittner, J. B., May, K., & Silver, N. C. (2003). A Monte Carlo evaluation of tests for comparing dependent correlations. *The Journal of General Psychology*, *130*(2), 149–168.

Hopcroft, J. E. & Ullman, J. D. (1979). *Introduction to Automata Theory, Languages and Computation*. Reading, MA: Addison-Wesley.

Hörnel, D. (1997). MELONET 1: Neural nets for inventing baroque-style chorale variations. In M. Jordan, M. Kearns, & S. Solla (Eds.), *Advances in Neural Information Processing 10*. Cambridge, MA: MIT Press.

Hörnel, D. & Olbrich, F. (1999). Comparative style analysis with neural networks. In *Proceedings of the 1999 International Computer Music Conference* (pp. 433–436). San Francisco: ICMA.

Howard, P. G. (1993). *The Design and Analysis of Efficient Lossless Data Compression Systems*. Doctoral dissertation, Department of Computer Science, Brown University, Providence, USA.

Huang, X., Alleva, F., Hon, H., Hwang, M., Lee, K., & Rosenfeld, R. (1993). The SPHINX-II speech recognition system: An overview. *Computer, Speech and Language*, *2*, 137–148.

Hughes, M. (1977). A quantitative analysis. In M. Yeston (Ed.), *Readings in Schenker Analysis and Other Approaches* chapter 8. New Haven, CT: Yale University Press.

Hume, D. (1965). Of the standard of taste. In D. Hume (Ed.), *Essays: Moral, Political and Literary* (pp. 231–255). Oxford: Oxford University Press.

Huron, D. (1997). *Humdrum* and *Kern*: Selective feature encoding. In E. Selfridge-Field (Ed.), *Beyond MIDI: The Handbook of Musical Codes* (pp. 375–401). Cambridge, MA: MIT Press.

Huron, D. (2001). Tone and voice: A derivation of the rules of voice-leading from perceptual principles. *Music Perception*, *19*(1), 1–64.

Jackendoff, R. (1987). *Consciousness and the Computational Mind*. Cambridge, MA: MIT Press.

Jackson, P. W. & Messick, S. (1965). The person, the product and the response: Conceptual problems in the assessment of creativity. *Journal of Personality*, *33*, 309–329.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge University Press.

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In W. W. Cohen & H. Hirsh (Eds.), *Proceedings of the Eleventh International Conference on Machine Learning* (pp. 121–129). San Francisco, CA: Morgan Kaufmann.

Johnson-Laird, P. N. (1983). *Mental Models*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (1991). Jazz improvisation: A theory at the computational level. In P. Howell, R. West, & I. Cross (Eds.), *Representing Musical Structure* (pp. 291–325). London: Academic Press.

Jones, M. R. (1981). Music as a stimulus for psychological motion: Part I. Some determinants of expectancies. *Psychomusicology*, *1*(2), 34–51.

Jones, M. R. (1982). Music as a stimulus for psychological motion: Part II. An expectancy model. *Psychomusicology*, *2*(1), 1–13.

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

Jones, M. R. (1990). Learning and the development of expectancies: An interactionist approach. *Psychomusicology*, *9*(2), 193–228.

Jones, M. R. & Boltz, M. G. (1989). Dynamic attending and responses to time. *Psychological Review*, *96*(3), 459–491.

Justus, T. & Hutsler, J. J. (2005). Fundamental issues in the evolutionary psychology of music: Assessing innateness and domain specificity. *Music Perception*, *23*(1), 1–27.

Kant, I. (1952). *The Critique of Judgement*. Oxford: Clarendon Press. trans. J. C. Meredith.

Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recogniser. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *35*(3), 400–401.

Kessler, E. J., Hansen, C., & Shepard, R. N. (1984). Tonal schemata in the perception of music in Bali and the West. *Music Perception*, *2*(2), 131–165.

Kippen, J. & Bel, B. (1992). Modelling music with grammars. In A. Marsden & A. Pople (Eds.), *Computer Representations and Models in Music* (pp. 207–238). London: Academic Press.

Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(3), 226–239.

Kneser, R. & Ney, H. (1995). Improved backing-off for *m*-gram language modelling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1 (pp. 181–184). Washington, DC: IEEE.

Knopoff, L. & Hutchinson, W. (1983). Entropy as a measure of style: The influence of sample length. *Journal of Music Theory*, *27*, 75–97.

Koelsch, S., Gunter, T., & Friederici, A. D. (2000). Brain indices of music processing: "Nonmusicians" are musical. *Journal of Cognitive Neuroscience*, *12*(3), 520–541.

Kohavi, R. (1995a). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, volume 2 (pp. 1137–1145). San Mateo, CA: Morgan Kaufmann.

Kohavi, R. (1995b). *Wrappers for Performance Enhancement and Oblivious Decision Graphs*. Doctoral dissertation, Department of Computer Science, Stanford University, USA.

Kohavi, R. & John, G. H. (1996). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1-2), 273–324.

Kratus, J. (1989). A time analysis of the compositional processes used by children aged 7 to 11. *Journal of Research in Music Education*, *37*, 5–20.

Kratus, J. (1994). Relationship among children's music audiation and their compositional processes and products. *Journal of Research in Music Education*, *42*(2), 115–130.

Krumhansl, C. L. (1979). The psychological representation of musical pitch in a tonal context. *Cognitive Psychology*, *11*, 346–374.

Krumhansl, C. L. (1990). *Cognitive Foundations of Musical Pitch*. Oxford: Oxford University Press.

Krumhansl, C. L. (1995a). Effects of musical context on similarity and expectancy. *Systematische Musikwissenschaft*, *3*(2), 211–250.

Krumhansl, C. L. (1995b). Music psychology and music theory: Problems and prospects. *Music Theory Spectrum*, *17*, 53–90.

Krumhansl, C. L. (1997). Effects of perceptual organisation and musical form on melodic expectancies. In M. Leman (Ed.), *Music, Gestalt and Computing: Studies in Cognitive Systematic Musicology* (pp. 294–319). Berlin: Springer.

Krumhansl, C. L. & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organisation in a spatial representation of musical keys. *Psychological Review*, *89*(4), 334–368.

Krumhansl, C. L., Louhivuori, J., Toiviainen, P., Järvinen, T., & Eerola, T. (1999). Melodic expectation in Finnish spiritual hymns: Convergence of statistical, behavioural and computational approaches. *Music Perception*, *17*(2), 151–195.

Krumhansl, C. L. & Shepard, R. N. (1979). Quantification of the hierarchy of tonal functions within a diatonic context. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(4), 579–594.

Krumhansl, C. L., Toivanen, P., Eerola, T., Toiviainen, P., Järvinen, T., & Louhivuori, J. (2000). Cross-cultural music cognition: Cognitive methodology applied to North Sami yoiks. *Cognition*, *76*(1), 13–58.

Krzanowski, W. J. (1988). *Principles of Multivariate Analysis*. Oxford, UK: Oxford University Press.

Kuhn, R. & De Mori, R. (1990). A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(6), 570–583.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: The University of Chicago Press.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the Growth of Knowledge* (pp. 91–196). Cambridge, UK: Cambridge University Press.

Large, E. W., Palmer, C., & Pollack, J. B. (1995). Reduced memory representations for music. *Cognitive Science*, *19*(1), 53–96.

Larsson, N. J. (1996). Extended application of suffix trees to data compression. In J. A. Storer & M. Cohn (Eds.), *Proceedings of the IEEE Data Compression Conference* (pp. 190–199). Washington, DC: IEEE Computer Society Press.

Lartillot, O., Dubnov, S., Assayag, G., & Bejerano, G. (2001). Automatic modelling of musical style. In *Proceedings of the 2001 International Computer Music Conference* (pp. 447–454). San Francisco: ICMA.

Lee, C. (1991). The perception of metrical structure: Experimental evidence and a model. In P. Howell, R. West, & I. Cross (Eds.), *Representing Musical Structure* (pp. 59–127). London: Academic Press.

Lerdahl, F. (1988a). Cognitive constraints on compositional systems. In J. A. Sloboda (Ed.), *Generative Processes in Music: The Psychology of Performance, Improvisation and Composition* (pp. 231–259). Oxford: Clarendon Press.

Lerdahl, F. (1988b). Tonal pitch space. *Music Perception*, *5*(3), 315–350.

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

Lewin, D. (1987). *Generalised Musical Intervals and Transformations*. New Haven/London: Yale University Press.

Liegeoise-Chauvel, C., Peretz, I., Babai, M., Laguitton, V., & Chauvel, P. (1998). Contribution of different cortical areas in the temporal lobes to music processing. *Brain*, *121*(10), 1853–1867.

Lomax, A. (1962). Song structure and social structure. *Ethnology*, *1*(4), 425–452.

Longuet-Higgins, H. C. (1962a). Letter to a musical friend. *The Music Review*, *23*, 244–248.

Longuet-Higgins, H. C. (1962b). Second letter to a musical friend. *The Music Review*, *23*, 271–280.

Longuet-Higgins, H. C. (1981). Artificial intelligence – a new theoretical psychology? *Cognition*, *10*(1–3), 197–200.

Longuet-Higgins, H. C. & Steedman, M. J. (1971). On interpreting Bach. In B. Meltzer & D. Michie (Eds.), *Machine Intelligence 6* (pp. 221–241). Edinburgh, UK: Edinburgh University Press.

MacKay, D. J. C. (1998). Introduction to Monte Carlo methods. In M. I. Jordan (Ed.), *Learning in Graphical Models*, NATO Science Series (pp. 175–204). Dordrecht, The Netherlands: Kluwer Academic Press.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge University Press.

Mahajan, M., Beeferman, D., & Huang, X. (1999). Improved topic-dependent language modeling using information retrieval techniques. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. Washington, DC: IEEE Computer Society Press.

Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Manzara, L. C., Witten, I. H., & James, M. (1992). On the entropy of music: An experiment with Bach chorale melodies. *Leonardo*, *2*(1), 81–88.

Marr, D. (1982). *Vision*. San Francisco: W. H. Freeman.

Marsden, A. (2000). Music, intelligence and artificiality. In E. R. Miranda (Ed.), *Readings in Music and Artificial Intelligence* (pp. 15–28). Amsterdam: Harwood Academic Publishers.

Martin, S., Hamacher, C., Liermann, J., Wessel, F., & Ney, H. (1999). Assessment of smoothing methods and complex stochastic language modeling. In *Proceedings of the Sixth European Conference on Speech Communication and Technology* (pp. 1939–1942). Budapest, Hungary: ISCA.

Mayer, R. E. (1999). Fifty years of creativity research. In R. J. Sternberg (Ed.), *Handbook of Creativity* (pp. 449–460). Cambridge, UK: Cambridge University Press.

McClamrock, R. (1991). Marr's three levels: A re–evaluation. *Minds and Machines*, *1*, 185–196.

McDermott, J. & Hauser, M. (2005). The origins of music: Innateness, uniqueness and evolution. *Music Perception*, *23*(1), 29–59.

Meredith, D. (1996). *The Logical Structure of an Algorithmic Theory of Tonal Music*. Doctoral dissertation, Faculty of Music, University of Oxford. Submitted.

Meredith, D. (2003). Pitch spelling algorithms. In R. Kopiez, A. C. Lehmann, I. Wolther, & C. Wolf (Eds.), *Proceedings of the Fifth Triennial ESCOM Conference* (pp. 204–207). Hanover, Germany: Institute for Research in Music Education.

Meyer, L. B. (1956). *Emotion and Meaning in Music*. Chicago: University of Chicago Press.

Meyer, L. B. (1957). Meaning in music and information theory. *Journal of Aesthetics and Art Criticism*, *15*(4), 412–424.

Meyer, L. B. (1967). *Music, the Arts and Ideas: Patterns and Predictions in Twentieth-century Music*. Chicago: University of Chicago Press.

Meyer, L. B. (1973). *Explaining Music: Essays and Explorations*. Chicago: University of Chicago Press.

Miller, L. K. (1987). Determinants of melody span in a developmentally disabled musical savant. *Psychology of Music*, *15*, 76–89.

Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw Hill.

Moffat, A. (1990). Implementing the PPM data compression scheme. *IEEE Transactions on Communications*, *38*(11), 1917–1921.

Moffat, A., Neal, R., & Witten, I. H. (1998). Arithmetic coding revisited. *ACM Transactions on Information Systems*, *16*(3), 256–294.

Moffat, A., Sharman, N., Witten, I. H., & Bell, T. C. (1994). An empirical evaluation of coding methods for multi-symbol alphabets. *Information Processing & Management*, *30*(6), 791–804.

Mozer, M. C. (1994). Neural network music composition by prediction: Exploring the benefits of psychoacoustic constraints and multi–scale processing. *Connection Science*, *6*(2–3), 247–280.

Narmour, E. (1990). *The Analysis and Cognition of Basic Melodic Structures: The Implication-realisation Model*. Chicago: University of Chicago Press.

Narmour, E. (1991). The top–down and bottom–up systems of musical implication: Building on Meyer's theory of emotional syntax. *Music Perception*, *9*(1), 1–26.

Narmour, E. (1992). *The Analysis and Cognition of Melodic Complexity: The Implication-realisation Model*. Chicago: University of Chicago Press.

Narmour, E. (1999). Hierarchical expectation and musical style. In D. Deutsch (Ed.), *The Psychology of Music* (Second ed.). (pp. 441–472). New York: Academic Press.

Newell, A. & Simon, H. A. (1976). Computer science as empirical enquiry: Symbols and search. *Communications of the ACM*, *19*(3), 113–126.

Nolan, D. (1997). Quantitative parsimony. *British Journal for the Philosophy of Science*, *48*(3), 329–343.

Oram, N. & Cuddy, L. L. (1995). Responsiveness of Western adults to pitch-distributional information in melodic sequences. *Psychological Research*, *57*(2), 103–118.

Page, M. P. A. (1999). Modelling the perception of musical sequences with self-organising neural networks. In N. Griffith & P. M. Todd (Eds.), *Musical Networks: Parallel Distributed Perception and Performance* (pp. 175–198). Cambridge, MA: MIT Press/Bradford Books.

Palmer, C. & Krumhansl, C. L. (1990). Mental representations for musical metre. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(4), 728–741.

Palmer, R. (Ed.). (1983). *Folk songs collected by Ralph Vaughan Williams*. London: Dent.

Papadopoulos, G. & Wiggins, G. A. (1999). AI methods for algorithmic composition: A survey, a critical view and future prospects. In *Proceedings of the AISB'99 Symposium on Musical Creativity* (pp. 110–117). Brighton, UK: SSAISB.

Paul, G. (1993). Approaches to abductive reasoning: An overview. *Artificial Intelligence Review*, *7*(2), 109–152.

Pearce, M. T. & Wiggins, G. A. (2001). Towards a framework for the evaluation of machine compositions. In *Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in the Arts and Sciences* (pp. 22–32). Brighton, UK: SSAISB.

Peretz, I. (1990). Processing of local and global musical information by unilateral brain-damaged patients. *Brain*, *113*(4), 1185–1205.

Phon-Amnuaisuk, S., Tuson, A., & Wiggins, G. A. (1999). Evolving musical harmonisation. In *Proceedings of the 1999 International Conference on Artificial Neural Networks and Genetic Algorithms*. New York: IEEE.

Pickens, J., Bello, J. P., Monti, G., Sandler, M. B., Crawford, T., Dovey, M., & Byrd, D. (2003). Polyphonic score retrieval using polyphonic audio queries: A harmonic modeling approach. *Journal of New Music Research*, *32*(2), 223–236.

Pinkerton, R. C. (1956). Information theory and melody. *Scientific American*, *194*(2), 77–86.

Plucker, J. A. & Renzulli, J. S. (1999). Psychometric approaches to the study of human creativity. In R. J. Sternberg (Ed.), *Handbook of Creativity* (pp. 35–61). Cambridge, UK: Cambridge University Press.

Plunkett, K., Karmiloff-Smith, A., Bates, E., Elman, J., & Johnson, M. H. (1997). Connectionism and developmental psychology. *Journal of Child Psychology and Psychiatry*, *38*(1), 53–80.

Plunkett, K. & Marchman, V. (1996). Learning from a connectionist model of the acquisition of the past tense. *Cognition*, *61*(3), 299–308.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, *46*(1), 77–105.

Ponsford, D., Wiggins, G. A., & Mellish, C. (1999). Statistical learning of harmonic movement. *Journal of New Music Research*, *28*(2), 150–177.

Popper, K. (1959). *The Logic of Scientific Discovery*. London: Hutchinson and Co.

Povel, D. J. (2004). *Melody Generator 7.0 User's Guide*. University of Nijmegen, The Netherlands: Nijmegen Institute for Cognition and Information. (Retrieved 25 November, 2004, from `http://www.socsci.kun.nl/~povel/Melody/UsersGuideMG7.0.pdf`).

Povel, D. J. & Essens, P. (1985). Perception of temporal patterns. *Music Perception*, *2*(4), 411–440.

Povel, D. J. & Jansen, E. (2002). Harmonic factors in the perception of tonal melodies. *Music Perception*, *20*(1), 51–85.

Povel, D. J. & Okkerman, H. (1981). Accents in equitone sequences. *Perception and Psychophysics*, *30*(6), 565–572.

Priest, T. (2001). Using creativity assessment experience to nurture and predict compositional creativity. *Journal of Research in Music Education*, *49*(3), 245–257.

Pylyshyn, Z. (1984). *Computation and Cognition*. Cambridge, MA: MIT Press.

Pylyshyn, Z. (1989). Computing in cognitive science. In M. I. Posner (Ed.), *Foundations of Cognitive Science* (pp. 51–91). Cambridge, MA: MIT Press.

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–285.

Reis, B. Y. (1999). *Simulating Music Learning with Autonomous Listening Agents: Entropy, Ambiguity and Context*. Doctoral dissertation, Computer Laboratory, University of Cambridge, UK.

Riemenschneider, A. (1941). *371 Harmonised Chorales and 69 Chorale Melodies with Figured Bass*. New York: G. Schirmer, Inc.

Ritchie, G. (2001). Assessing creativity. In *Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in the Arts and Sciences* (pp. 3–11). Brighton, UK: SSAISB.

Roads, C. (1985a). Grammars as representations for music. In C. Roads & J. Strawn (Eds.), *Foundations of Computer Music* (pp. 403–442). Cambridge, MA: MIT Press.

Roads, C. (1985b). Research in music and artificial intelligence. *ACM Computing Surveys*, *17*(2), 163–190.

Ron, D., Singer, Y., & Tishby, N. (1996). The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning*, *25*(2–3), 117–149.

Rosner, B. S. & Meyer, L. B. (1982). Melodic processes and the perception of music. In D. Deutsch (Ed.), *The Psychology of Music* (pp. 317–341). New York: Academic Press.

Rosner, B. S. & Meyer, L. B. (1986). The perceptual roles of melodic process, contour and form. *Music Perception*, *4*(1), 1–40.

Rothstein, J. (1992). *MIDI: A Comprehensive Introduction*. Oxford: Oxford University Press.

Rowe, R. J. (1992). Machine composing and listening with Cypher. *Computer Music Journal*, *16*(1), 43–63.

Rumelhart, D. E., Hinton, G., & Williams, R. (1986). Learning internal representations through error propagation. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing: Experiments in the Microstructure of Cognition*, volume 1 (pp. 25–40). Cambridge, MA: MIT Press.

Russo, F. A. & Cuddy, L. L. (1999). A common origin for vocal accuracy and melodic expectancy: Vocal constraints. Paper presented at the Joint Meeting of the Acoustical Society of America and the European Acoustics Association. Published in Journal of the Acoustical Society of America, 105, 1217.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52.

Schaffrath, H. (1992). The ESAC databases and MAPPET software. *Computing in Musicology*, *8*, 66.

Schaffrath, H. (1994). The ESAC electronic songbooks. *Computing in Musicology*, *9*, 78.

Schaffrath, H. (1995). The Essen folksong collection. In D. Huron (Ed.), *Database containing 6,255 folksong transcriptions in the Kern format and a 34-page research guide [computer database]*. Menlo Park, CA: CCARH.

Schellenberg, E. G. (1996). Expectancy in melody: Tests of the implication-realisation model. *Cognition*, *58*(1), 75–125.

Schellenberg, E. G. (1997). Simplifying the implication-realisation model of melodic expectancy. *Music Perception*, *14*(3), 295–318.

Schellenberg, E. G., Adachi, M., Purdy, K. T., & McKinnon, M. C. (2002). Expectancy in melody: Tests of children and adults. *Journal of Experimental Psychology: General*, *131*(4), 511–537.

Schellenberg, E. G. & Trehub, S. E. (2003). Good pitch memory is widespread. *Psychological Science*, *14*(3), 262–266.

Schmuckler, M. A. (1989). Expectation in music: Investigation of melodic and harmonic processes. *Music Perception*, *7*(2), 109–150.

Schmuckler, M. A. (1990). The performance of global expectations. *Psychomusicology*, *9*(2), 122–147.

Schmuckler, M. A. (1997). Expectancy effects in memory for melodies. *Canadian Journal of Experimental Psychology*, *51*(4), 292–305.

Schmuckler, M. A. & Boltz, M. G. (1994). Harmonic and rhythmic influences on musical expectancy. *Perception and Psychophysics*, *56*(3), 313–325.

Schubert, E. (2001). Continuous measurement of self-report emotional responses to music. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and Emotion* (pp. 393–414). Oxford: Oxford University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423 and 623–656.

Sharp, C. J. (Ed.). (1920). *English folk songs*, volume 1-2, selected edition. London: Novello.

Shepard, R. N. (1982). Structural representations of musical pitch. In D. Deutsch (Ed.), *Psychology of Music* (pp. 343–390). New York: Academic Press.

Simon, H. A. (1973). The structure of ill–structured problems. *Artificial Intelligence*, *4*, 181–201.

Simon, H. A. & Kaplan, C. A. (1989). Foundations of cognitive science. In M. I. Posner (Ed.), *Foundations of Cognitive Science* (pp. 1–47). Cambridge, MA: MIT Press.

Simons, M., Ney, H., & Martin, S. (1997). Distant bigram language modeling using maximum entropy. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (pp. 787–790). Washington, DC: IEEE Computer Society Press.

Simonton, D. K. (1998). Arnheim award address to division 10 of the American Psychological Association. *Creativity Research Journal*, *11*(2), 103–110.

Sloboda, J. (1985). *The Musical Mind: The Cognitive Psychology of Music*. Oxford: Oxford Science Press.

Smaill, A., Wiggins, G. A., & Harris, M. (1993). Hierarchical music representation for composition and analysis. *Computers and the Humanities*, *27*, 7–17.

Sober, E. (1981). The principle of parsimony. *British Journal for the Philosophy of Science*, *32*(2), 145–156.

Steedman, M. (1984). A generative grammar for jazz chord sequences. *Music Perception*, *2*(1), 52–77.

Steedman, M. (1996). The blues and the abstract truth: Music and mental models. In A. Garnham & J. Oakhill (Eds.), *Mental Models in Cognitive Science* (pp. 305–318). Mahwah, NJ: Erlbaum.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251.

Stobart, H. & Cross, I. (2000). The Andean anacrusis? Rhythmic structure and perception in Easter songs of Northern Potosí, Bolivia. *British Journal of Ethnomusicology*, *9*(2), 63–94.

Sun, R. & Giles, C. L. (2001). Sequence learning: From recognition and prediction to sequenctial decision making. *IEEE Intelligent Systems*, *16*(4), 67–70.

Sundberg, J. & Lindblom, B. (1976). Generative theories in language and music descriptions. *Cognition*, *4*(1), 99–122.

Sundberg, J. & Lindblom, B. (1991). Generative theories for describing musical structure. In P. Howell, R. West, & I. Cross (Eds.), *Representing Musical Structure* (pp. 245–272). London: Academic Press.

Tax, D. M. J., van Breukelen, M., Duin, R. P. W., & Kittler, J. (2000). Combining multiple classifiers by averaging or by multiplying? *Pattern Recognition*, *33*(99), 1475–1485.

Teahan, W. J. (1998). *Modelling English Text*. Doctoral dissertation, Department of Computer Science, University of Waikato, Hamilton, New Zealand.

Teahan, W. J. & Cleary, J. G. (1996). The entropy of English using PPM-based models. In J. A. Storer & M. Cohn (Eds.), *Proceedings of the IEEE Data Compression Conference*. Washington, DC: IEEE Computer Society Press.

Teahan, W. J. & Cleary, J. G. (1997). Models of English text. In J. A. Storer & M. Cohn (Eds.), *Proceedings of the IEEE Data Compression Conference*. Washington, DC: IEEE Computer Society Press.

Tekman, H. G. (1997). Interactions of perceived intensity, duration and pitch in pure tone sequences. *Music Perception*, *14*(3), 281–294.

Temperley, D. (1999). What's key for key? The Krumhansl-Schmuckler key-finding algorithm reconsidered. *Music Perception*, *17*(1), 65–100.

Temperley, D. (2001). *The Cognition of Basic Musical Structures*. Cambridge, MA: MIT Press.

Temperley, D. (2003). Communicative pressure and the evolution of musical styles. *Music Perception*, *21*(3), 313–337.

Thompson, W. F. (1993). Modelling perceived relationships between melody, harmony and key. *Perception and Psychophysics*, *53*(1), 13–24.

Thompson, W. F. (1994). Sensitivity to combinations of musical parameters: Pitch with duration and pitch pattern with durational pattern. *Perception and Psychophysics*, *56*(3), 363–374.

Thompson, W. F. (1996). Eugene Narmour: *The analysis and cognition of basic musical structures* (1990) and *The analysis and cognition of melodic complexity* (1992): A review and empirical assessment. *Journal of the American Musicological Society*, *49*(1), 127–145.

Thompson, W. F., Cuddy, L. L., & Plaus, C. (1997). Expectancies generated by melodic intervals: Evaluation of principles of melodic implication in a melody-completion task. *Perception and Psychophysics*, *59*(7), 1069–1076.

Thompson, W. F. & Stainton, M. (1996). Using *Humdrum* to analyse melodic structure: An assessment of Narmour's implication-realisation model. *Computing in Musicology*, *12*, 24–33.

Thompson, W. F. & Stainton, M. (1998). Expectancy in Bohemian folk song melodies: Evaluation of implicative principles for implicative and closural intervals. *Music Perception*, *15*(3), 231–252.

Todd, N. P. M., O'Boyle, D. J., & Lee, C. S. (1999). A sensory-motor theory of rhythm, time perception and beat induction. *Journal of New Music Research*, *28*(1), 5–28.

Toiviainen, P. (2000). Symbolic AI versus connecionism in music research. In E. R. Miranda (Ed.), *Readings in Music and Artificial Intelligence* (pp. 47–68). Amsterdam: Harwood Academic Publishers.

Toiviainen, P. & Eerola, T. (2004). The role of accent periodicities in metre induction: A classification study. In S. D. Lipscomb, R. Ashley, R. O. Gjerdingen, & P. Webster (Eds.), *Proceedings of the Eighth International Conference of Music Perception and Cognition* (pp. 422–425). Adelaide, Australia: Causal Productions.

Toiviainen, P. & Krumhansl, C. L. (2003). Measuring and modelling real-time responses to music: The dynamics of tonality induction. *Perception*, *32*(6), 741–766.

Trehub, S. E. (1999). Human processing predispositions and musical universals. In N. L. Wallin, B. Merker, & S. Brown (Eds.), *The Origins of Music* (pp. 427–448). Cambridge, MA: MIT Press.

Triviño-Rodriguez, J. L. & Morales-Bueno, R. (2001). Using multi-attribute prediction suffix graphs to predict and generate music. *Computer Music Journal*, *25*(3), 62–79.

Ukkonen, E. (1995). On-line construction of suffix trees. *Algorithmica*, *14*(3), 249–260.

Unyk, A. M. & Carlsen, J. C. (1987). The influence of expectancy on melodic perception. *Psychomusicology*, *7*(1), 3–23.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer.

von Hippel, P. T. (2000a). Questioning a melodic archetype: Do listeners use gap-fill to classify melodies? *Music Perception*, *18*(2), 139–153.

von Hippel, P. T. (2000b). Redefining pitch proximity: Tessitura and mobility as constraints on melodic intervals. *Music Perception*, *17*(3), 315–127.

von Hippel, P. T. (2002). Melodic-expectation rules as learned heuristics. In C. Stevens, D. Burnham, E. Schubert, & J. Renwick (Eds.), *Proceedings of the Seventh International Conference on Music Perception and Cognition* (pp. 315–317). Adelaide, Australia: Causal Productions.

von Hippel, P. T. & Huron, D. (2000). Why do skips precede reversals? The effects of tessitura on melodic structure. *Music Perception*, *18*(1), 59–85.

Vos, P. G. (2000). Tonality induction: Theoretical problems and dilemmas. *Music Perception*, *17*(4), 403–416.

Vos, P. G. & Pasveer, D. (2002). Goodness ratings of melodic openings and closures. *Perception and Psychophysics*, *64*(4), 631–639.

Vos, P. G. & Troost, J. M. (1989). Ascending and descending melodic intervals: Statistical findings and their perceptual relevance. *Music Perception*, *6*(4), 383–396.

Šerman, M. & Griffith, N. J. L. (2004). Investigating melodic segmentation through the temporal multi-scaling framework. *Musicæ Scientiæ*, *Special Tenth Anniversary issue on musical creativity*.

Webster, P. (1987). Conceptual bases for creative thinking in music. In J. C. Peery & I. W. Peery (Eds.), *Music and Child Development* (pp. 158–174). New York: Springer.

Webster, P. & Hickey, M. (1995). Rating scales and their use in assessing children's music compositions. *The Quarterly Journal of Music Teaching and Learning*, *VI*(4), 28–44.

West, R., Howell, P., & Cross, I. (1985). Modelling perceived musical structure. In P. Howell, I. Cross, & R. West (Eds.), *Musical Structure and Cognition* (pp. 21–52). London: Academic Press Inc.

Westhead, M. D. & Smaill, A. (1993). Automatic characterisation of musical style. In M. Smith, A. Smaill, & G. Wiggins (Eds.), *Music Education: An Artificial Intelligence Approach* (pp. 157–170). Berlin: Springer.

Wiggins, G. A. (1998). Music, syntax, and the meaning of "meaning". In *Proceedings of the First Symposium on Music and Computers* (pp. 18–23). Corfu, Greece: Ionian University.

Wiggins, G. A., Harris, M., & Smaill, A. (1989). Representing music for analysis and composition. In M. Balaban, K. Ebcioğlu, O. Laske, C. Lischka, & L. Soriso (Eds.), *Proceedings of the Second Workshop on AI and Music* (pp. 63–71). Menlo Park, CA: AAAI.

Wiggins, G. A., Miranda, E., Smaill, A., & Harris, M. (1993). A framework for the evaluation of music representation systems. *Computer Music Journal*, *17*(3), 31–42.

Wiggins, G. A. & Smaill, A. (2000). What can artificial intelligence bring to the musician? In E. R. Miranda (Ed.), *Readings in Music and Artificial Intelligence* (pp. 29–46). Amsterdam: Harwood Academic Publishers.

Willems, F. M. J., Shtarkov, Y. M., & Tjalkens, T. J. (1995). The context-tree weighting method: Basic properties. *IEEE Transactions on Information Theory*, *41*(3), 653–664.

Witten, I. H. & Bell, T. C. (1991). The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, *37*(4), 1085–1094.

Witten, I. H., Manzara, L. C., & Conklin, D. (1994). Comparing human and computational models of music prediction. *Computer Music Journal*, *18*(1), 70–80.

Witten, I. H., Neal, R. M., & Cleary, J. G. (1987). Arithmetic coding for data compression. *Communications of the ACM*, *30*(6), 520–541.

Xu, L., Krzyzak, A., & Suen, C. Y. (1992). Methods of combining multiple classifiers and thier applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, *22*(3), 418–435.

Youngblood, J. E. (1958). Style as information. *Journal of Music Theory*, *2*, 24–35.

Ziv, J. & Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, *24*(5), 530–536.