Chapter 1

Computational Modelling of Music Cognition and Musical Creativity

Geraint A. Wiggins, Marcus T. Pearce and Daniel Müllensiefen
Centre for Cognition, Computation and Culture
Goldsmiths, University of London

1.1 Introduction

This chapter is about computational modelling of the process of musical composition, based on a cognitive model of human behaviour. The idea is to try to study not only the requirements for a computer system which is capable of musical composition, but also to relate it to human behaviour during the same process, so that it may, perhaps, work in the same way as a human composer, but also so that it may, more likely, help us understand how human composers work. Pearce et al. (2002) give a fuller discussion of the motivations behind this endeavour.

We take a purist approach to our modelling: we are aiming, ultimately, at a computer system which we can claim to be creative. Therefore, we must address in advance the criticism that usually arises in these circumstances: "a computer can't be creative because it can only do what it has explicitly been programmed to do". This argument does not hold, because, with the advent of machine learning, it is no longer true that a computer is limited to what its programmer explicitly tells it, especially in a relatively unsupervised learning task like composition (as compared with the usually-supervised task of learning, say, the piano). Thus, a creative system based on machine learning can, in principle, be given credit for creative output, much as Wolfgang Amadeus Mozart is deemed the creator of the Magic Flute, and not Leopold Mozart, Wolfgang's father, teacher and *de facto* agent.

Because music is a very complex phenomenon, we focus on a relatively[1] simple aspect, which is relatively easy to isolate from the many other aspects of music: tonal melody. In order to compose music, one normally needs to learn about it by hearing it, so we begin with a perceptual model, which has proven capable of simulating relevant aspects of human listening behaviour better than any other in the literature. We also consider the application of this model to a different task, musical phrase segmentation, because doing so adds weight to its status as a good, if preliminary, model of human cognition. We then consider using this model to generate tonal melodies, and show how one might go about

evaluating the resulting model of composition scientifically. Before we can begin this discussion, we will need to cover some background material, and introduce some descriptive tools, which are the subject of the next section.

1.2 Background

1.2.1 Introduction

In this section, we explain the basis of our approach to the cognitive modelling of musical creativity and supply background material to the various detailed sections to follow. We begin by motivating cognitive modelling itself, and then argue why doing so is relevant to the study of musical behaviour. We make a distinction between different kinds of cognitive model, which serve different purposes in the context of research. Next, we outline an approach to modelling creative behaviour, within which we frame our discussion. Finally, we briefly survey the literature in modelling of music cognition and musical composition, and in the evaluation of creative behaviour, to supply background for the later presentation.

1.2.2 Methodology

Our starting point: Cognitive modelling

Cognitive science as a research field dates back to the 1950s and '60s. It arises from a view of the brain as an information-processing machine, and the mind as an epiphenomenon arising in turn from that processing. The aim is to understand the operation of the mind and brain at various interconnected levels of abstraction, in the expectation that, ultimately, cognitive scientists will be able to explain the operation of both mind and brain, from the level of physical neurons up to the level of consciousness. There is an important distinction between the study of the *operation of the mind in general* and the *resulting emergent behaviour of particular minds or groups of minds*; the former is our focus here. This cognitive focus follows, in the current context, from a view of music, not as a Romantic, quasi-Platonic and transcendent entity with an absolute definition in the external world, but as a fundamentally social phenomenon, driven by and formed from the human urge to communicate. Only thus can we account for the multifarious musics of the human world and the way they change over time, given the lack of any strong evidence for the existence of innate specifically musical abilities shaped directly by evolution (Justus and Hutsler, 2005). Necessarily, therefore, we look for the source of music in humanity, and also in individual (but not particular) humans, the latter being our main interest here.

The difficulty with studying minds and brains is that they are very difficult to measure. The only way one can measure a mind is by recording its effect on the world and therefore one can only infer the causes of one's results. Brains are a little more accessible, but ethical issues restricts the extent to which we

can do controlled experiments with them[2], and anyway they are so complicated that we lack the technology to study them in the detail we really need. To overcome these problems, cognitive scientists have tended to focus on particular aspects of measurable behaviour, in an abstract way, ignoring surrounding detail, in the hope of understanding them in isolation before moving on to more inclusive theories. The choice of abstraction is crucial, because, done wrongly, it can obscure parts of the phenomenon being studied or blur the distinctions between different effects.

Computational cognitive models

With the advent of computers, cognitive scientists were able to take their models of the mind to new levels of precision. Previously, they were able to describe what effects arose from what stimulus, but it was impossible to give a mechanistic theory from which predictions could be made, because doing so would have been an intractable pen-and-paper exercise, enormously time-consuming and error-prone. However, with fast computers and access to large, high-quality databases of stimuli, it is now possible to embody a cognitive theory as a computer program, and thus apply it to large amounts of data, and to test its consequences exhaustively—thus, importantly, generating new hypotheses for testing against human behaviour from these predictions. In a cyclic way, we can then refine our theory to account for incorrect predictions, and try again. In addition to goodness of fit to the observed data and behaviours, we prefer simpler models over more complex ones, models that selectively predict just the observed data, and, finally, models that generate surprising, but true, predictions (Cutting et al., 1992, Honing, 2007).

As well as supplying a way forward, computational modelling gives cognitive scientists a new and useful challenge: to define their working abstraction and their theory precisely enough that it can be given an operational interpretation as a computer program. Much research in computer representation of music is also engaged in this challenge (e.g., Marsden, 2000, Wiggins et al., 1993).

Another issue which is brought into sharp focus is the distinction between modelling *what a phenomenon does*, and modelling *how it does it*, which have been labelled *descriptive* and *explanatory* modelling (Wiggins, 2007); Marr (1982) and McClamrock (1991) also discuss these and related issues. To understand this distinction, an analogy is helpful. Consider models of the weather. Such a model could be made by taking a barometer, correlating atmospheric pressure with the weather and the wind direction, and writing down whether it was raining or not at the time. Given enough data, this simple look-up table will probably predict the weather correctly much of the time. However, it only computes its predictions in terms of observed correlated connections: it encodes nothing of the mechanisms by which the weather operates, and therefore cannot explain how the weather works; nor can it account for conditions it has not met before, unless by some naïve generalisation such as interpolation. If it is indeed reasonably accurate, the model can nevertheless be

useful predictor of the weather, and so we might say it *describes* the weather to some degree. Now imagine a different super-computer-based weather model, which has detailed information about the same empirical data, but which also encodes knowledge of physics (for example, of the process whereby liquid water will precipitate from humid air as temperature drops). This physical model need only be in terms of mathematical equations such as Boyle's Law, and not, for example, in terms of the movement of individual molecules of gas, but it nevertheless captures a different *kind* of detail from the descriptive one above— and we can ask it "why?". So this *explanatory* model gives an account of the weather by saying (at another level of abstraction) *how* the effects described by both models actually arise. Like the descriptive model, we can test it by giving it conditions that we have newly experienced for the first time, and checking its predictions against reality, and if they turn out to be wrong, one source of potential error is now the mechanism itself.

A final useful concept here will be that of the *meta-model*, named from the Greek $\mu\varepsilon\tau\alpha$, meaning *after* or *beyond,* as in "metaphysics". We use this to refer to the use of a model, intended for and validated with respect to a particular (cognitive) phenomenon, (directly or indirectly) to predict the behaviour of another related but different phenomenon for which it was neither intended nor designed. It is useful to make this distinction because this capacity adds considerable weight to the argument that the model is in some sense a *good* model in general terms (Honing, 2007), and can also support a case that it is an explanatory one. We give an example of such a model and a meta-model derived from it below.

Computational cognitive modelling of creative behaviour

Since the point of this chapter is to consider creative applications of cognitive models, we need a framework within which to do so. Boden (1990) proposes a model of creative behaviour which revolves around the notion of a *conceptual space* and its exploration by creative agents. The conceptual space is a set of *concepts* which are deemed to be acceptable as examples of whatever is being created. Implicitly, the conceptual space may include partially defined concepts too. *Exploratory creativity* is the process of exploring a given conceptual space; *transformational creativity* is the process of changing the rules which delimit the conceptual space. Boden (1998) also makes an important distinction between mere membership of a conceptual space and the *value* of a member of the space, which is extrinsically defined, but not precisely. Various other models of creativity exist (e.g., Koestler, 1964, Wallas, 1926), but are not sufficiently detailed for implementation; Ritchie (2007) gives an alternative view of ways to study creative systems, but it does not suit our purposes here.

Boden's model, however, is amenable to implementation. Wiggins (2006a,b) provides one possible formalisation, presenting a Creative Systems Framework (CSF) which may be directly constructed, or used to identify aspects of creative systems and compare them with each other and with human behaviour. There

is not space to present the full framework here; it suffices to echo Boden's idea of a conceptual space, defined by a mutable rule set, R, and a further mutable set of rules, E, according to which the quality of the items created can be evaluated. This dichotomy is important—for example, it is possible to recognise a joke without thinking it to be a good one—and so it is necessary to separate these things. An explicit component of Wiggins' formalism which is only implicit in Boden's original thought experiment is the idea of a *traversal strategy*, T, which is used by a creative agent to explore the conceptual space—in other words, while it is actually doing the creative stuff. This is necessary for a computer system (otherwise nothing will happen!), but also for an explicit model of a specific creative agent: the difference, for example, between a first year music student and an experienced professional organist harmonising a chorale melody lies not just in the quality of the output produced, but also in the encoding of the strategies used: the unexceptional student is likely to use trial and error to some extent, whereas the organist can intuitively "see" the right harmonisation.

Throughout the rest of this chapter, we use all three of the concepts behind the abstract rule sets outlined above, referring to them as R, T, and E, to identify as precisely as we can which aspect of a creative system we are discussing.

1.2.3 "Non-cognitive" musical composition systems

For completeness, we must acknowledge the existence of a substantial body of work in autonomous systems for musical composition which is not directly related to music cognition and therefore not directly related to this chapter. The earliest such system of which we are aware is that of Hiller and Isaacson (1959), where a stochastic model was used to generate a musical score, which was subsequently performed by human musicians. Since the 1950s, various attempts have been made at creating music, without explicit reference to the processes humans use in doing so. In many of these attempts, the emphasis is on reproducing the style of existing (or formerly existing) composers. In context of the CSF (see above), the focus is then primarily on R and E; T is usually treated mainly as an implementation detail, without regard to simulation of human behaviour. A particularly good example of this approach is CHORAL (Ebcioğlu, 1988), a rule-based expert system implemented in a specially-written Backtracking Specification Language (BSL) and used for the harmonisation of chorale melodies in the style of J. S. Bach. Here, R and E are intertwined in the code of the program and it is not clear how to decide which is which (though there is an excellent specification of several hundred Bach-harmonic rules in Ebcioğlu's thesis, which may well be a good approximation to R); T is entirely implicit, and is obscured to most readers because it is encoded in Ebcioğlu's backtracking strategy.

Other systems make more of an explicit attempt to model evaluation on the basis of musical attributes perceived by a hypothetical listener. For example,

Robertson et al. (1998) present HERMAN, a system which is capable of generating continuous music, whose emotional property can be varied from neutral to "scary". Rutherford and Wiggins (2002) demonstrate empirically that human responses do to an extent match the intention of the program's operator. The focus here was again on R and E, though the difference between them was made more explicit by the use of specific heuristics; T was again relegated to a matter of implementation.

It is important to understand that both CHORAL and HERMAN, and many other systems like them, rely on music theory for the basis of their operation, and, as such, encode those aspects of music cognition which are implicit in music theory (which, we suggest, are many). However, it is difficult to argue that such knowledge-based systems actually model human creative behaviour, because they are programmed entities, and merely do what their programmers have made them do: in the terms outlined above, they are descriptive, and not explanatory, models. We suggest that, for an autonomous composition system to be considered genuinely "creative", it is necessary (though not sufficient) that the system include a significant element of *autonomous learning*. Then, while the *urge* to create may well be instilled by a programmer, the products of creativity are not.

No review of this field would be complete without mention of the work of David Cope (2005). Cope's *Experiments in Musical Intelligence* program has long standing in computer music research as the standard of stylistic re-composition. Regrettably, however, the publications do not tell us how it works, and therefore it is impossible to discuss its behaviour in detail.

1.2.4 Non-computational models of music cognition

There is a long history of efforts to develop models of listening-related music cognition that are both formal (although not specifically computational) and general. From the point of view of the CSF, we view all these theories as contributing primarily to R, in a hypothetical creative system, though E is presumably affected too, since these theories are not about creating music, but listening to it.

Perhaps the earliest attempt was that of Simon and Sumner (1968), who assume that music cognition involves pattern induction and attempt to define a formal language for describing the patterns perceived and used by humans in processing musical sequences. They begin with the notion of an *alphabet*, an ordered set of symbols, for representing the range of possible values for a particular musical dimension (e.g., melody, harmony, rhythm and form, using alphabets for diatonic notes, triads, duration, stress and formal structure). Simon and Sumner define three kinds of operation. First, subset operations may be defined to derive more abstract alphabets from existing ones. Second, sequences of symbols may be described by patterns of operations that relate a symbol to its predecessor (e.g., *same* or *next*). Finally, a pattern of operations may be replaced by an abstract symbol. According to this model, when we

listen to music, we first induce an alphabet, initial symbol and pattern consistent with what we hear and then use that pattern to extrapolate the sequence.

Deutsch and Feroe (1981) extend the pattern language of Simon and Sumner and flesh out its formal specification. They use it to define various common collections of notes (such as scales, triads and chords) through the recursive application of different operators to an alphabet based on the chromatic scale. Arguing that patterns are learnt through long-term exposure to a particular music style, they motivate their approach by appealing to parsimony of encoding (reduced representational redundancy) and constraints on memory and processing (through chunking). However, experiments have yielded mixed support for the predictions of the model (Boltz and Jones, 1986, Deutsch, 1980).

The *Generative Theory of Tonal Music* (GTTM) of Lerdahl and Jackendoff is probably the best-known effort to develop a comprehensive method for the structural description of tonal music. Inspired by the use of Chomskian grammars to describe language, the theory is intended to yield a hierarchical, structural description of any piece of Western tonal music, corresponding 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: *grouping structure*, the segmentation of the musical surface into units (e.g., motives, phrases); *metrical structure*, the pattern of periodically recurring strong and weak beats; *time-span reduction*, the relative structural importance of pitch events within contextually established rhythmic units; and *prolongational reduction*, 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, *preference 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. Thus, 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 and Krumhansl, 1990).

Narmour presents the *Implication-Realisation* (IR) theory of music cognition which, like GTTM, is intended to be general (although the initial presentation was restricted to melody) but which, in contrast to GTTMs static approach, starts with the dynamic processes involved in perceiving music in time. The theory posits two distinct perceptual systems: the *bottom-up* system is held to be hard-wired, innate and universal while the *top-down system* is held to be learnt through musical experience. The two-systems may conflict and, in any given situation, one may over-ride the implications generated by the other.

In the bottom-up system, sequences of melodic intervals vary in the degree of *closure* that they convey. 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, termed the *realised interval*. The expectations generated by implicative intervals for realised intervals are described by Narmour in terms of several principles of continuation which are influenced by the Gestalt principles of proximity, similarity, and good continuation. The IR model also specifies how the basic melodic structures combine together to form longer and more complex structural patterns of melodic implication within the IR theory. In particular, structures associated with weak closure may be *chained* to subsequent structures. In addition, 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.

The IR theory has inspired many quantitative implementations of its principles and a large body of experimental research testing its predictions as a theory of melodic expectation (Cuddy and Lunny, 1995, Krumhansl, 1995a, 1995b, Krumhansl et al., 2000, Schellenberg, 1996, 1997, Thompson et al., 1997), which we may interpret as supportive. Our experiments seem to add further support, but with the proviso that our system is (in Narmour's terms) top-down only (Pearce and Wiggins, 2006).

1.2.5 Computational models of music cognition

Given that we wish to base our autonomous creative system on behaviour that is learnt, rather than programmed, we need to identify a starting point, from which the learnt behaviour can arise. In humans, this starting point seems to be the ability to hear music and perceive its internal structure; it is hard to imagine how musically creative behaviour could arise otherwise, unless it is an intrinsic property of human brains. There is no evidence for this latter claim, but there is evidence that music is learnt: without learning, it is very hard to account for the ubiquity of music in human society while still explaining the variety of musics in different cultures and sub-cultures. Various authors (e.g., Bown and Wiggins, 2008, Cross, 2007, Justus and Hutsler, 2005, Mithen, 2006) have studied these questions; the consensus seems to be that music cognition and music creation co-evolve—and, indeed, we arguably see the cultural aspects of this process continuing in the present day, and not only in (pre)history.

There are not very many computational models of music cognition in the literature, and those that do exist span a wide range of musical dimensions—music cognition is too complicated a phenomenon to be modelled directly all in one go. Aspects of the general frameworks described above have been implemented piecemeal. The approach usually taken is the standard scientific reductionist approach: attempt to understand each aspect of the problem while holding the others fixed, then try to understand their interactions, and only subsequently to put all the understanding together. Again a general distinction can be made between programmed-rule-based and machine learning approaches. It is worth mentioning that this distinction is independent of the structural divisions of the CSF: each of the three rule sets, R, T and E, may be either human-programmed or learned, so long as there is a context in which they can interact. However, there is a difference in relation to the descriptive/exploratory distinction: a pre-programmed rule-based system (of which our simple weather-predictor, above, was an extreme example), is less likely to be an explanatory model than a descriptive one, because, by definition, it does not give any account of how the rules arise in context of their eventual usage.

On the machine learning side, Jamshed Bharucha developed a connectionist model of harmony based on a sequential feed-forward neural network. 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 and Stoeckig, 1986, 1987). In addition, the network model learned 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. Large et al. (1995) examined the ability of another neural network architecture, RAAM (Pollack, 1990), to acquire reduced representations of Western children's melodies represented as tree structures according to music-theoretic predictions (Lerdahl and Jackendoff, 1983). 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 certainty with which the trained network reconstructed events correlated well with cognitive representations of structural importance as assessed by empirical data on the events retained by trained pianists across improvised variations on the melodies.

Perhaps the most complete computational theory to date is that of Temperley (2001), which is inspired to an extent by GTTM. Temperley proposed preference rule models of a range of fundamental processes in music cognition which include metre recognition, melodic segmentation, voice separation in polyphonic music, pitch spelling, chord analysis, and key identification. The rule models reflect sophisticated knowledge from music theory and are implemented in a suite of analysis tools named *Melisma* whose source code is publicly available. When applied to real-world analysis problems the *Melisma* tools

generally exhibit reasonable performance (see below regarding melodic segmentation or Meredith, 2006, regarding pitch spelling) and in some areas have become a standard for rule-based music analysis algorithms. Most of the algorithmic models bear little underlying conceptual coherence and make strong use of domain-specific knowledge as reflected by the respective rules and their combination. Temperley (2007) aims at a reformulation of some of these rule-based models in the general probabilistic framework of Bayesian statistics. He derives a so-called *pitch* and a *rhythm model* based on frequency counts in different music corpora and applies them to several musical processes such as metre-determination, key-finding and melodic error detection.

As the Bayesian models do not always outperform the rule-based algorithms, the value of the Bayesian reformulation seems to lie rather in the more coherent underlying theory, although a more comprehensive and rigorous evaluation is still required (Pearce et al., 2007).

1.2.6 Computational cognitive models of musical composition

By comparison with cognitive-scientific research on musical listening, cognitive processes in composition remain largely unexamined (Baroni, 1999, Sloboda, 1985). This section reviews research on the cognitive modelling of music composition with an emphasis on computational approaches; Deliège and Wiggins (2006) present a range of work in the non-computational context.

Johnson-Laird (1991) argues that it is fundamental to understand what the mind has to compute in order to generate an acceptable jazz improvisation before examining the precise nature of the algorithms by which it does so.[3] To study the intrinsic constraints of the task, Johnson-Laird applied grammars of different expressive powers to different subcomponents of the problem. His results 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 a context free grammar.

Lerdahl (1988) explores the relationship between listening and composition and outlines some cognitive constraints that it places on the cognitive processes of composition. He 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 a *listening grammar* that, in turn, yields a structural description of the composition as perceived. 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 the two 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. He goes on to outline some constraints, largely based on the preference rules and stability conditions of GTTM, placed on compositional grammars by this need to recover the intended structural organisation from the musical surface by the listening grammar.

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

Baroni (1999) discusses grammars for modelling the cognitive processes involved in musical listening and composition, basing his arguments on his own grammars for the structural analysis of a number of musical repertoires (Baroni et al., 1992). He 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, he 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, he 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 top-down.

Perhaps surprisingly, given that we are now considering composition, in terms of the CSF, the emphasis is, again, on R and maybe E: the method of traversal, T, being treated almost as a given.

Turning now to machine-learning approaches, Conklin (2003) examines four methods of generating high-probability music according to a statistical model. The simplest is sequential random sampling: an event is sampled from the estimated event distribution at each sequential position up to a given length. Events are generated in a random walk, so there is a danger of straying into local minima in the space of possible compositions. Even so, most statistical generation of music uses this method.

The Hidden Markov Model (HMM) addresses these problems; it generates observed events from hidden states (Rabiner, 1989). An HMM is trained by 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. This can be achieved efficiently for a first-order HMM using the Viterbi algorithm; a similar algorithm exists for first-order (visible) Markov models. However, Viterbi's running time increases exponentially in the context length of the underlying Markov model (Conklin, 2003), which means that it is not good for practical use. However, there do exist tractable methods for sampling from complex statistical models (such as those presented here) which address the limitations of random sampling (Conklin, 2003). We return to this below.

Notwithstanding the problems of complexity, Conklin makes a real attempt to address the question of what a traversal strategy, T, might be, as well as considering R carefully. In our work, we follow his statistical approach, but note that something more semantically sensitive will be appropriate in future work.

1.2.7 Evaluation of creative behaviour

The evaluation of creative behaviour, either within a creative system, or from outside it, is very difficult because of the subjectivity involved, and because individual outputs cannot necessarily be said to be representative of the system's capability.

On the computational side, *analysis by synthesis* has been used to evaluate computational models of composition by generating pieces and evaluating them with respect to the objectives of the implemented model. The method has a long history; Ames and Domino (1992) argue that a primary advantage of computational analysis of musical style is the ability to evaluate new pieces generated from an implemented theory. However, evaluation of the generated music raises methodological issues which have typically compromised the potential benefits thus afforded (Pearce et al., 2002). Often, compositions are evaluated with a single subjective comment, e.g., : "[the compositions] are realistic enough that an unknowing listener cannot discern their artificial origin" (Ames and Domino, 1992, pp. 186). This lack of precision makes it hard to compare theories intersubjectively.

Other research has used expert stylistic analyses to evaluate computer compositions. This is possible when a computational model is developed to account for some reasonably well-defined stylistic competence or according to criteria derived from music theory or music psychology. For example, Ponsford et al. (1999) gave an informal stylistic appraisal of the harmonic progressions generated by their *n*-gram models.

However, even when stylistic analyses are undertaken by groups of experts, the results obtained are typically still qualitative. For fully intersubjective analysis by synthesis, the evaluation of the generated compositions must be

empirical. One could use an adaptation of the Turing test, where subjects are presented with pairs of compositions (one computer-generated, the other human-composed) and asked which they believe to be the computer-generated one (Marsden, 2000). Musical Turing tests yield empirical, quantitative results which may be appraised intersubjectively and have demonstrated the inability of subjects to distinguish reliably between computer- and human-composed music. But the method suffers from three major difficulties: it can be biased by preconceptions about computer music, allows ill-informed judgements, and fails to examine the criteria being used to judge the compositions.

Assessing human creativity is no easier, but at least one technique has been proposed that seems promising. Amabile (1996) proposes a conceptual definition of creativity in terms of processes resulting in novel, appropriate solutions to heuristic, open-ended or ill-defined tasks. However, while agreeing that creativity can only be assessed through subjective assessments of products, she criticises the use of *a priori* theoretical definitions of creativity in rating schemes and failure to distinguish creativity from other constructs. While a conceptual definition is important for guiding empirical research, a clear operational definition is necessary for the development of useful empirical methods of assessment. Accordingly, she presents a consensual definition of creativity in 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 internally consistent (independent judges agree in their ratings of creativity), one can empirically examine the objective or subjective features of creative products which contribute to their perceived creativity.

Amabile used this operational definition to develop the *consensual assessment technique* (CAT), an empirical method for evaluating creativity. Its requirements are that the task be open-ended enough to permit considerable flexibility and novelty in the response, which must be an observable product which can be rated by judges. Regarding the procedure, the judges must:

1. be experienced in the relevant domain;
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 stimuli;
5. be presented with stimuli and provide ratings in orders randomised differently for each judge.

Most importantly, in analysing the collected data, the inter-judge reliability of the subjective rating scales must be determined. If—and only if—reliability is high, we may correlate creativity ratings with other objective or subjective features of creative products.

Numerous 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 (Amabile, 1996, ch. 3, gives a review).

The CAT overcomes the limitations of the Turing test in evaluating computational models of musical composition. First, it requires the use of human judges expert in the task domain. Second, since it has been developed for research on human creativity, no mention is made of the computational origins of the stimuli; this avoids bias due to preconceptions. Third, and most importantly, the methodology allows more detailed examination of the objective and subjective dimensions of the creative products. Crucially, the objective attributes of the products may include features of the generative models (corresponding with cognitive or stylistic hypotheses) which produced them. Thus, we can empirically compare different musicological theories of a given style or hypotheses about the cognitive processes involved in composing in that style.

We propose to use the CAT in evaluating creative computer systems as well as human ones.

1.3 Towards a computational model of music cognition

1.3.1 Introduction

Having laid out the background of our approach, and supplied a context of extant research, we now present our model of melody cognition (the *Information Dynamics Of Music* or IDyOM model). We describe it in three parts: first, the computational model itself; second, its application to melodic pitch expectation; and third, the extension of the same model to melodic grouping (which we then call a meta-model of music cognition). As with the other cognitive models described above, we will view the expectation model as supplying R, and perhaps some of E, in our creative system. For want of literature on the subject, we begin with a naïve statistical T.

The following is a brief summary; detailed presentations of the model are available elsewhere (Pearce, 2005, Pearce et al., 2005, Pearce and Wiggins, 2004).

1.3.2 The Computational Model

The representation scheme

We use a *multiple viewpoint system* (Conklin and Witten, 1995) as the basis of our representation scheme. The scheme takes as its *musical surface* (Jackendoff, 1987) sequences of note events, representing the instantiation of a finite number of discrete features or attributes. An event consists of a number of *basic features* representing its onset time, duration, pitch and so on. Basic features are associated with an alphabet: a finite set of symbols determining the

possible instantiations of that feature in a concrete note.

The representation scheme also allows for the construction of *derived features* which can be computed from the values of one or more basic features (e.g., inter-onset interval, pitch interval, contour, and scale degree). In some locations in a melody, a given derived feature may be undefined. Furthermore, it is possible to define derived features that represent attributes of non-adjacent notes and compound features may be defined to represent interactions between primitive features.

To ensure that our results pertain to real-world musical phenomena, and to ensure ecological validity, we use music data from existing repertoires of music. Here, we use data derived from scores but the representation scheme is rather flexible and could be extended to represent expressive aspects of music performance (e.g., dynamics, expressive timing). Although we focus on melody, and not all musics have an equivalent analogue of the Western notion, stream segregation (Bregman, 1990) appears to be a basic perceptual process. Furthermore, the multiple viewpoints framework has been extended to accommodate the representation of homophonic and polyphonic music (Conklin, 2002).

The modelling strategy

IDyOM itself is based on *n*-gram models commonly used in statistical language modelling (Manning and Schütze, 1999). An *n*-gram is a sequence of *n* symbols and an *n*-gram model is simply a collection of such sequences each of which is associated with a frequency count. During the *training* of the statistical model, these counts are acquired through an analysis of some corpus of sequences (the training set) in the target domain. When the trained model is exposed to a sequence drawn from the target domain, it uses the frequency counts associated with *n*-grams to estimate a probability distribution governing the identity of the next symbol in the sequence given the $n-1$ preceding symbols. The quantity $n-1$ is known as the *order* of the model and represents the number of symbols making up the context within which a prediction is made.

The modelling process begins by choosing a set of basic features that we are interested in predicting. As these basic features are treated as independent attributes, their probabilities are computed separately and in turn, and the probability of a note is simply the product of the probabilities of its attributes. Here we consider the example of predicting pitch alone.

The most elementary *n*-gram model of melodic pitch structure (a monogram model where $n=1$) simply tabulates the frequency of occurrence for each chromatic pitch encountered in a traversal of each melody in the training set. During prediction, the expectations of the model are governed by a zeroth-order pitch distribution derived from the frequency counts and do not depend on the preceding context of the melody. In a digram model (where $n=2$), however, frequency counts are maintained for sequences of two pitch symbols and predictions are governed by a first-order pitch distribution derived from the

frequency counts associated with only those digrams whose initial pitch symbol matches the final pitch symbol in the melodic context.

Fixed order models such as these suffer from a number of problems. Low-order models (such as the monogram model discussed above) clearly fail to provide an adequate account of the structural influence of the context on expectations. However, increasing the order can prevent the model from capturing much of the statistical regularity present in the training set. An extreme case occurs when the model encounters an n-gram that does not appear in the training set in which case it returns an estimated probability of zero. In order to address these problems, the IDyOM model maintains frequency counts during training for n-grams of all possible values of n in any given context. During prediction, distributions are estimated using a weighted sum of all models below a variable order bound. This bound is determined in each predictive context using simple heuristics designed to minimise uncertainty. The combination is designed such that higher-order predictions (which are more specific to the context) receive greater weighting than lower-order predictions (which are more general). In a given melodic context, therefore, the predictions of the model may reflect the influence of both the digram model and (to a lesser extent) the monogram model discussed above. Furthermore, in addition to the general, low-order statistical regularities captured by these two models, the predictions of the IDyOM model can also reflect higher-order regularities which are even more specific to the current melodic context (to the extent that these exist in the training set).

Inference over multiple features:

One final issue to be covered regards the manner in which IDyOM exploits the representation of multiple features of the musical surface described above. The modelling process begins with the selection, by hand, of a set of features of interest and the training of distinct n-gram models for each of these features. For each note in a melody, each feature is predicted using two models: first, the *long-term* model that was trained over the entire training set in the previous step; and second, a *short-term* model that is trained incrementally for each individual melody being predicted. Figure 1.1 illustrates this (and other aspects of the model of pitch expectation and the meta-model of melodic segmentation, to which we return below).
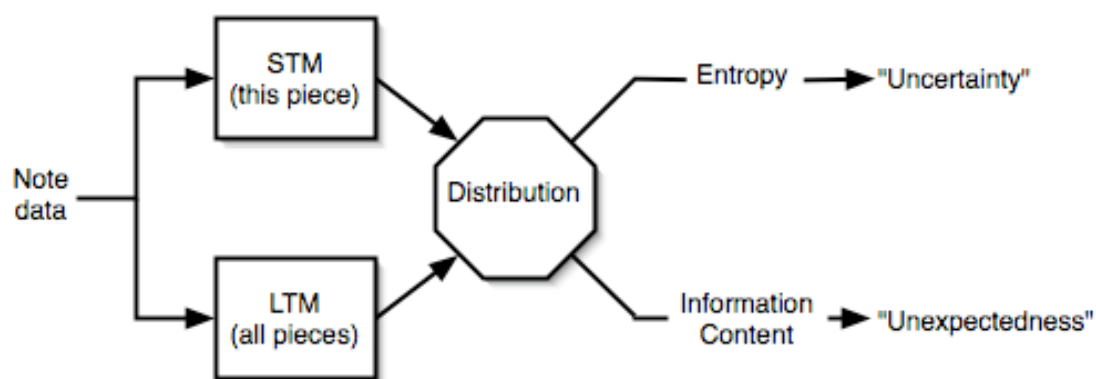
Figure 1.1: Our development of Pearce's (2005) cognitive model. The central distribution gives the model of pitch expectation; the entropy and information content outputs, coupled with some peak-picking, constitute the meta-model of melodic segmentation.

The task of combining the predictions from all these models is achieved in two stages both of which use a weighted multiplicative combination scheme in which greater weights are assigned to models whose predictions are associated with lower entropy (or uncertainty) at that point in the melody. In this scheme, a combined distribution is achieved by taking the product of the weighted probability estimates returned by each model for each possible value of the pitch of the next note and then normalising such that the combined estimates sum to unity over the pitch alphabet. The entropy-based weighting method and the use of a multiplicative as opposed to a additive combination scheme both improve the performance of the model in predicting the pitches of unseen melodies (Pearce et al., 2005, Pearce and Wiggins, 2004).

In the first stage of model combination, the predictions of models for different features are combined for the long-term and short-term models separately. Distributions from models of derived features are first converted into distributions over the alphabet of the basic feature from which they are derived (e.g., in order to combine a distribution over pitch contours with one over scale degrees, first we need to convert both into distributions over chromatic pitch). If any feature (derived or otherwise) is undefined at a given location in a melody, a model of that feature will not contribute to the predictions of the overall system at that location. In the second stage, the two combined distributions (long-term and short-term) resulting from the first step are combined into a single distribution which represents the overall system's final expectations regarding the pitch of the next note in the melody. The use of long- and short-term models is intended to reflect the influences on expectation of both existing extra-opus and incrementally increasing intra-opus knowledge while the use of multiple features is intended to reflect the influence of regularities in many dimensions of the musical surface.

1.3.3 Modelling melodic pitch expectancy

The conditional probabilities output by IDyOM in a given melodic context may be interpreted as contextual expectations about the nature of the forthcoming note. Pearce and Wiggins (2006) compare the melodic pitch expectations of the model with those of listeners in the context of single intervals (Cuddy and Lunny, 1995), at particular points in British folk songs (Schellenberg, 1996) and throughout two chorale melodies (Manzara et al., 1992). The results demonstrate that the statistical system predicts the expectations of listeners as least as well as the two-factor model of Schellenberg (1997) and significantly better in the case of more complex melodic contexts.

1.3.4 Modelling melodic segmentation

Musical segmentation is a fundamental process in music-cognitive theory and simulation (e.g., Cambouropoulos, 1996, Lerdahl and Jackendoff, 1983, Potter et al., 2007, Wiggins, 2007). In this section, we show how our model of pitch expectation can be used to predict human judgements of melodic segment boundaries. Inevitably, this meta-model (see section 1.2.2) is not superior to all existing segmentation models from the literature because it includes no direct encoding of the musical features that we know determine segmentation: metrical structure, harmony, and so on. However, it performs surprisingly well, in comparison with other descriptive, programmed models. Our model can predict both large-scale and small-scale boundaries in music.

From a musicological perspective, it has been proposed that perceptual groups are associated with points of closure where the ongoing cognitive process of expectation is disrupted either because the context fails to stimulate strong expectations for any particular continuation or because the actual continuation is unexpected (Meyer, 1957, Narmour, 1990). In addition, empirical psychological research has demonstrated that infants and adults use the implicitly learnt statistical properties of pitch (Saffran et al., 1990), pitch interval (Saffran and Griepentrog, 2001) and scale degree (Saffran, 2003) sequences to identify segment boundaries on the basis of higher digram ($n=2$) transition probabilities within than between groups. Finally, in machine learning and computational linguistics, algorithms based on the idea of segmenting before unexpected events perform reasonably well in identifying word boundaries in infant-directed speech (Brent, 1999, Cohen et al., 2007, Elman, 1990). There is some evidence that high predictive uncertainty is also associated with word boundaries (Cohen et al., 2007).

Drawing on this background, we achieve our meta-model by applying information-theoretic principles (Shannon, 1948) to the distributions produced by the statistical model of melodic expectation. In particular, we represent unexpectedness of a note by its *information content* (the negative log of its conditional probability); and we represent uncertainty about the identity of the next note by the *entropy* (the average information content) of the distribution

governing the note's identity computed from both the STM and LTM. Our prediction of large-scale structure works by looking for relatively large, simultaneous, positive change-points in the entropy and information content of the music at each note, as it proceeds (Potter et al., 2007), essentially implementing Narmour's (1990) proposal. Entropy measures uncertainty, so a downward trend followed by a sudden increase corresponds with closure at the end of a phrase and then the relative uncertainty generated by not knowing what comes next. Unexpectness measures something related (though of course the beginning of each new phrase need not necessarily be very unexpected in absolute terms!).

Here we focus on the effects of unexpectedness (modelled by information content) on low-level melodic segmentation, leaving the role of entropy for future research. Using a model of both pitch and rhythmic structure (inter-onset interval and rests), we derive an information-content profile for the notes in a melody, from which we identify segment boundaries by picking peaks at points where the note-by-note information content is high *relative to the local context* (see Müllensiefen et al., 2008, for further details).

The performance of our model was evaluated in two studies where we compared its prediction accuracy to the performance of several other models specifically designed for melodic segmentation, such as *Grouper* (Temperley, 2001), the *LBDM* (Cambouropoulos, 2001), and three of the *Grouping Preference Rules* from GTTM (Lerdahl and Jackendoff, 1983).

The data for the first evaluation study was collected from 25 expert judges in an experimental setting. Their task was to indicate phrase endings within each of 15 popular melodies on a score through repeated listenings to each melody. For all the 1250 note events in this dataset we computed the F-score, a widely used evaluation measure in information retrieval (see e.g., Jurafsky and Martin, 2000), to indicate the correspondence between the algorithmic segmentation solutions and the boundaries selected by at least 50% of the experimental participants; this is merely a majority vote—more principled approaches will be applied in future work. The F score can vary between 0, when there is no correspondence, and 1, indicating perfect agreement between model and ground truth data.

The second evaluation study used 1705 German folk songs from the Essen collection (Schaffrath, 1995) as ground truth data. This dataset comprised 78,995 notes at an average of about 46 events per melody and overall about 12% of notes fall before boundaries. Phrase boundaries were annotated by the musicologist who encoded the data, and the F-scores on this dataset reflect the degree to which the boundaries predicted by a given model correspond to those annotated in the scores.

Model	F (Essen data)	F (experimental data)
Grouper	0.65	0.76
LBDM	0.62	0.71
GPR2a	0.60	0.76
IDyOM	0.58	0.77
GPR2b	0.38	0.16
GPR3a	0.34	0.22
GPR3d	0.29	0.08
Always	0.22	0.14
Never	0.00	0.00

Table 1.1: Segmentation model performances (F-score) on the Essen folksong data and data from an experimental study.

As can be seen from Table 1.1, the IDyOM model reaches a performance comparable to the three best performing segmentation models, namely Grouper, LBDM, and the GPR2a rule from GTTM. An acceptable melodic segmentation can be obtained by picking the peaks in the information content profile produced by the general-purpose learning model, suggesting that unexpectedness (as measured by information content) is very strongly related to boundary detection in melodies.

1.3.5 Discussion

In this section, we have described in outline the IDyOM model of music cognition that has been developed over several years, and which is still developing. While this model is still in its infancy, it has demonstrated a capacity to perform reasonably reliably and to a reasonable level of competence on certain well-defined, restricted, appropriate tasks. What is more, the model is able to describe two distinct (though not separate) phenomena, pitch expectation and phrase segmentation; we believe this property is methodologically important enough to deserve a name, and so we call the phrase segmentation model, which is derived from the pitch expectation model without changing it, a *meta-model* (because it is a model *beyond* the original model using the same process). Finally, the IDyOM model provides a mechanism learned from, but independent from any prior knowledge of, the data over which it operates and is therefore a candidate explanatory model of why the corresponding observed human behaviour is as it is, at this level of abstraction. IDyOM provides supporting evidence, therefore, for the hypothesis that, at some level of abstraction, this kind of music-cognitive processing is effectively modelled by statistics and probability (Huron, 2006, Temperley, 2007).

How, then, does this model relate to the generation of music? In terms of the

CSF, we can use IDyOM to estimate the probability of an entire melodic composition, giving us R, and choose a threshold below which any composition will be deemed "not a melody". So much is easy; defining T and E, however, is less so. First, what makes a "good" melody? Ponsford et al. (1999) initially hypothesised that music with high probability would be good, but this was quickly shown not to be the case: rather, very highly probable pieces tend to be syntactically correct, but musically dull. We suggest, in fact, that one way to characterise E would be to look at dynamic changes in the information-theoretic measures introduced above, as they vary with time. However, before we can properly consider trying to find *good* solutions, we need to be able to find any solution which conforms to R to any reasonable level. Minimally, this is what T is for. In the next section, we present the first stages of development on a potential creative system based on the cognitive model described above, which works by applying a standard method of string generation to the production of melodies from its learnt Markov model.

Of course, the model we have presented here leaves a lot to be desired in terms of general musical properties, such as those in harmony, rhythm, metre, or in musical contexts where these notions do not hold sway or are, at least, different. This is the way of the reductionist scientific method: one must delay gratification for quite a long time, when modelling complex phenomena.

However, we believe that our approach is applicable in the above contexts, either in addition to or instead of the existing data and representation: our underlying learning model is amenable to study in any musical context representable by discrete symbols. In principle, those symbols can denote anything from complex polyphony to harmonic labels, to sophisticated representations of timbre and sound spectra. Even in continuous contexts, there are related models which may apply, though there are open questions as how to combine them with the discrete ones. Since human perception tends to be categorical, we may even hypothesise that the continuous models are more than we need—at least until we have empirical evidence otherwise.

The immediate focus required, therefore, to expand this thinking to other styles and kinds of music, is on representations at an appropriate level of abstraction of the percepts generated by our senses, which can then be used in supporting or falsifying our learning model. Evidently, there is a rich vein of work to be mined here, and some of it will be future work for us.

1.4 A simple computational model of musical creativity

1.4.1 Introduction

We now describe our experimental exploration of the generative capacity of our perceptual model. Following Johnson-Laird (1991), we analyse the computational constraints of the melody composition task in two ways: first, examining whether our learned finite context grammars can compose stylistically-successful melodies or whether more expressive grammars are

needed; and second, determining which representational structures are needed for the composition of successful melodies.

Our experiment is designed to test the hypothesis that our statistical models are capable of generating melodies which are deemed stylistically successful in the context of a specified tradition. Three multiple-feature systems (Pearce, 2005) were trained on a dataset of chorale melodies, and were then used to generate melodies which were empirically evaluated. As described below, the three IDyOM systems are equivalent except for the sets of derived features they use to generate their pitch predictions.

Our work differs in several ways from extant statistical modelling for music generation, in particular, in that no symbolic constraints were imposed on the generation process—it was based entirely on the learnt models. This focuses the analysis more sharply on the inherent capacities of statistical finite context grammars, since our goal was to examine the synthetic capabilities of purely statistical, data-driven models of melodic structure. While most previous approaches used sequential random sampling to generate music from statistical models, to generate our output, we used the *Metropolis-Hastings algorithm*, a Markov Chain Monte Carlo (MCMC) sampling method (MacKay, 1998). The following description applies it within our generation framework. Given a trained multiple-feature model m for some basic feature τ_b, in order to sample from the target distribution $p_m (s \in [\tau_b]^*)$, the algorithm constructs a Markov chain in the space of possible feature sequences $[\tau_b]^*$ as follows (Notation: '←' denotes assignment; '++' denotes incrementation; 'an event' means a musical note, which is a collection of features, in the terms used above):

1. number of iterations $N \leftarrow$ a large value; iteration number $k \leftarrow 0$; initial state s_0 some feature sequence $t^j_1 \in [\tau_b]^*$ of length j;

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

3. let s'_k be the sequence obtained by replacing event t_i at index i of s_k with a new event t'_k sampled from a distribution q which may depend on the current state s_k – in the present context, an obvious choice for q would be $\{p_m (s (t \mid t_1^{i-2}) \} t \in [\tau_b]$;

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, $s_{k+1} \leftarrow s'_k$, else $s_{k+1} \leftarrow s_k$;

6. if $k < N$, k++ and iterate from 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 method of assessing the convergence of MCMCs nor of estimating the number of iterations required to obtain an unbiased sample (MacKay, 1998). Because these sampling algorithms explore the state space using a random walk, they can still be trapped in local minima. However, we expect that this method will be better than sequential random sampling at generating melodies that faithfully represent the inherent capacities of the Systems.[4]

Finally, to evaluate the systems as computational models of melodic composition, we developed a method based on the CAT. The method, described fully by Pearce (2005), obtains ratings by expert judges of the stylistic success of computer generated compositions and existing compositions in the target genre. The empirical nature of this method makes it preferable to the exclusively qualitative analyses typically adopted and we expect it to yield more revealing results than the Turing test methodology used in previous research (Hall and Smith, 1996, Triviño-Rodriguez and Morales-Bueno, 2001).

1.4.2 Hypotheses

We use three different Systems to examine which representational structures are needed for competent melody generation. Our null hypotheses are that each System can generate melodies rated as equally stylistically successful in the target style as existing, human-composed melodies.

System A is a single-feature system that generates predictions based on chromatic pitch alone. We expect the null hypothesis for the simplistic System A to be refuted. System B is multiple-feature system whose feature-set was optimised through forward, stepwise feature selection to provide the closest fit to human expectancy judgements in chorale melodies (Manzara et al., 1992). The set of three features selected included features related to tonality and melodic structure with an influence of both rhythmic structure and phrase structure. For this System, Baroni's (1999) proposal that composition and listening involve equivalent grammatical structures is relevant. If the representational structures underlying listening to and composition of music are similar, we would expect grammars which model perceptual processes well to generate satisfactory compositions. Since System B represents a satisfactory model of the cognition of pitch structure in the chorale genre, we may expect to retain the null hypothesis for this system.

System C is a multiple-feature system whose feature-set was optimised through forward, stepwise feature selection to yield the best pitch prediction performance over the entire chorale dataset. The selected set of nine features included features related to pitch, melodic structure and tonal structure with strong interactions with rhythmic, metric and phrase structure. In terms of model selection for music generation, highly predictive theories of a musical style, as measured by information content, should generate original and acceptable works in the style (Conklin and Witten, 1995). Systems A, B and C in turn exhibit increasing accuracy in predicting unseen melodies from the dataset. On

this basis, we may expect to retain the null hypothesis for System C.

1.4.3 Method

Our judges were 16 music researchers or students at City University, London, Goldsmiths, University of London, and the Royal College of Music. Seven judges reported high familiarity with the chorale genre and nine were moderately familiar.

Our dataset is a subset of the chorale melodies placed in the soprano voice and harmonised in four parts by J. S. Bach. These melodies are characterised by stepwise patterns of conjunct intervallic motion and simple, uniform rhythmic and metric structure. Phrase structure is explicitly notated. Most phrases begin on the tonic, mediant or dominant and end on the tonic or dominant; the final phrase almost always ends with a cadence to the tonic. Our stimuli were as follows. Seven existing *base* melodies were randomly selected from the set of chorales in the midrange of the distribution of average information content (cross-entropy) values computed by System A. All 7 were in common time; 6 were in major keys and 1 was minor; they were 8–14 bars (mean 11.14) and 33–57 events (mean 43.43) long. The base melodies were removed from the training dataset. 7 novel melodies were generated by each System, *via* 5000 iterations of Metropolis sampling using the 7 base chorales as initial states. Only pitch was sampled: time and key signatures and rhythmic and phrase structure were left unchanged. Figure 1.2 shows one base chorale melody and the three melodies generated using it; Pearce (2005) gives further examples.

Figure 1.2: An example of one base chorale melody and the three melodies generated using it.

Our judges supplied their responses individually and received instructions verbally and in writing. We told them they would hear a series of chorale

melodies in the style of Lutheran hymns and asked them to listen to each entire melody before answering two questions about it by placing circles on discrete scales in the response booklet. The first question was, "How successful is the composition as a chorale melody?" Judges were advised that their answers should reflect such factors as conformity to important stylistic features, tonal organisation, melodic shape and interval structure; and melodic form. Answers to this question were given on a seven-point numerical scale, 1–7, with anchors marked low (1), medium (4) and high (7). To promote an analytic approach to the task, judges were asked to briefly justify their responses to the first question. The second question was, "Do you recognise the melody?" Judges were advised to answer "yes" only if they could specifically identify the composition as one they were familiar with.

The experiment began with a practice session during which judges heard two human-composed melodies from the same original genre (but not one of those in the test set). These practice trials were intended to set a judgemental standard for the subsequent test session. This departs from the CAT, which encourages judges to rate each stimulus in relation to the others by experiencing all stimuli before making their ratings. However, here, we intended the judges to use their expertise to rate the stimuli against an absolute standard: 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; he then left the room before the start of the test session.

In the test session, the 28 melodies were presented to the judges, who responded 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 were presented sequentially. A reverse counterbalanced design was used, with eight of the judges listening to the melodies in one such order and the other eight listening to them in the reverse order.

1.4.4 Results

We report analyses of the 28 melodies from our test session: we discarded the data from the practice block.

Inter-judge Consistency

All but two of the 120 pairwise correlations between judges were significant at $p<0.05$ with a mean coefficient of $r(26)=0.65$ ($p<0.01$). Since there was no apparent reason to reject the judges involved in the two non-significant correlations, we did not do so. This high consistency warrants averaging the ratings for each stimulus across individual judges in subsequent analyses.

Presentation Order and Prior Familiarity

Two factors which might influence the judges' ratings are the order of presentation of the stimuli and prior familiarity. The correlation between the mean success ratings for judges in the two groups was $r(26)=0.91, p<0.01$ indicating a high degree of consistency across the two orders of presentation, and warranting the averaging of responses across the two groups; and, although the mean success ratings tended to be slightly higher when judges recognised the stimulus, a paired t test revealed no significant difference: $t(6)=2.07, p=0.08$.

Influence of Generative System and Base Chorale

Now we examine the primary question: the influence of generative system on the ratings of stylistic success. The mean success ratings for each stimulus, shown in Table 1.2a, suggest that the original chorale melodies were rated higher 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 an exception, attracting high average ratings of success. We analysed the data with Friedman's rank sum tests, using within-subjects factors for generative system with 4 levels (Original by J. S. Bach, System A, B, C) and base chorale with 7 levels (249, 238, 365, 264, 44, 153 and 147)

(a)

Base	A	B	C	Original	Mean
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
Mean	2.90	2.49	3.17	6.02	3.65

(b)

Statistic	A	B	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 1.2: (a) The mean success ratings for each stimulus and means aggregated by generative system and base chorale. (b) The median, quartiles and inter-quartile range of the mean success ratings for each generative

system.

We examined the influence of generative system in an unreplicated complete block 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 1.2b. The Friedman test revealed a significant within-subject effect of generative system on the mean success ratings: $\chi^2(3) = 33.4$, $p < 0.01$. We compared the factor levels pairwise using Wilcoxon rank sum tests with Holm's Bonferroni correction for multiple comparisons: the ratings for the original chorale melodies differ significantly from the ratings of melodies generated by all three computational systems ($p < 0.01$). Furthermore, the mean success ratings for the melodies generated by System B were found to be significantly different from those of the melodies generated by Systems A and C ($p < 0.03$). These results suggest that none of the systems is capable of consistently generating chorale melodies that are rated as equally stylistically successful as those in the dataset and that System B performed especially poorly.

1.4.5 Learning from Qualitative Feedback

Objective Features of the Chorales

Next, we aim to identify how the Systems lack compositionally, by examining which objective musical features of the stimuli the judges used in making their ratings of stylistic success. To achieve this, we analysed the stimuli qualitatively and developed a set of corresponding objective descriptors, which we then applied in a series of multiple regression analyses using the rating scheme, averaged across stimuli, as a dependent variable. We now present the descriptive variables, their quantitative coding and the analysis results.

The chorales generated by our systems are mostly not very stylistically characteristic of the dataset, especially in higher-level form. From the judges' qualitative comments, we identified stylistic constraints describing the stimuli and distinguishing the original melodies. We grouped them into five categories—pitch range; melodic structure; tonal structure; phrase structure; and rhythmic structure—each covered by one or more predictor variables.

Pitch Range

The dataset melodies span a pitch range of about an octave above and below C, favouring the centre of this range. The generated melodies are constrained to this range, but some tend towards extreme tessitura. We developed a predictor variable *pitch centre* to capture this difference, reflecting the absolute distance, in semitones, of the mean pitch of a melody from the mean pitch of the dataset (von Hippel, 2000). Another issue is the overall pitch range of the generated chorales. The dataset melodies span an average range of 11.8 semitones. By contrast, several of the generated melodies span pitch ranges of

16 or 17 semitones, with a mean pitch range of 13.9 semitones; others have a rather narrow pitch range. We captured these qualitative considerations in a quantitative predictor variable *pitch range*, representing the absolute distance, in semitones, of the pitch range of a melody from the mean pitch range of the dataset.

Melodic Structure

There are several ways in which the generated melodies do not consistently reproduce salient melodic features of the original chorales. The most obvious is a failure to maintain a stepwise pattern of movement. While some generated melodies are relatively coherent, others contain stylistically uncharacteristic leaps of an octave or more. Of 9042 intervals in the dataset melodies, only 57 exceed a perfect fifth and none exceeds an octave. To capture these deviations, we created a quantitative predictor variable called *interval size*, representing the number of intervals greater than a perfect octave in a melody. The generated chorales also contain uncharacteristic discords such as tritones or sevenths. Only 8 of the 9042 intervals in the dataset are tritones or sevenths (or their enharmonic equivalents). To capture these deviations, we created a quantitative predictor variable *interval dissonance*, representing the number of dissonant intervals greater than a perfect fourth in a melody.

Tonal Structure

Since System A operates exclusively over representations of pitch, it is not surprising that most of its melodies fail to establish a key note and exhibit little tonal structure. However, we might expect Systems B and C to do better. While the comments of the judges suggest otherwise, they may have arrived at a tonal interpretation at odds with the intended key of the base chorale. To independently estimate the perceived tonality of the test melodies, Krumhansl's (1990) key-finding algorithm, using the revised key profiles of Temperley (1999) was applied to each of the stimuli. The algorithm assigns the correct keys to all seven original chorale melodies. While the suggested keys of the melodies generated by System A confirm 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 System C, deviations from the base chorale key tend to be to related keys (either in the circle of fifths or through relative and parallel major/minor relationships). This suggests some success on the part of the more sophisticated systems in retaining the tonal characteristics of the base chorales.

Nonetheless, the generated melodies are often unacceptably chromatic, which obscures the tonality. Therefore, we developed a quantitative predictor called *chromaticism*, representing the number of chromatic tones in the algorithm's suggested key.

Phrase Structure

The generated chorales typically fail to reproduce the implied harmonic rhythm of the originals and its characteristically strong relationship to phrase structure. In particular, while some of the generated melodies close on the tonic, many fail to imply stylistically satisfactory harmonic closure. To capture such effects, we created a variable called *harmonic closure*, which is 0 if a melody closes on the tonic of the key assigned by the algorithm and 1 otherwise. Secondly, the generated melodies frequently fail to respect thematic repetition and development of melodic material embedded in the phrase structure of the chorales. However, these kinds of repetition and development of melodic material are not represented in the present model. Instead, as a simple indicator of complexity in phrase structure, we created a variable *phrase length*, which is 0 if all phrases are of equal length and 1 otherwise.

Rhythmic Structure

Although the chorale melodies in the dataset tend to be rhythmically simple, the judges' comments revealed that they were taking account of rhythmic structure. Therefore, we adapted three further quantitative predictors modelling rhythmic features from Eerola and North's (2000) expectancy-based model of melodic complexity. *Rhythmic density* is the mean number of events per tactus beat. *Rhythmic variability* is the degree of change in note duration (i.e., the standard deviation of the log of the event durations) in a melody. *Syncopation* estimates the degree of syncopation by assigning notes a strength in a metric hierarchy and averaging the strengths of all the notes in a melody; pulses are coded such that lower values are assigned to tones on metrically stronger beats. All three quantities increase the difficulty of perceiving or producing melodies (Eerola and North, 2000).

The mean success ratings for each stimulus were regressed on the predictor variables in a multiple regression analysis. Due to significant collinearity between the predictors, in each analysis, redundant predictors were removed through backwards stepwise elimination using the Akaike Information Criterion (Venables and Ripley, 2002).

Predictor	β	Std. Error	t	p
(Intercept)	6.4239	0.3912	16.42	0.0000
Pitch Range	-0.29	0.08	-3.57	< 0.01
Pitch Centre	-0.21	0.10	-2.01	< 0.1
Interval Dissonance	-0.70	0.28	-2.54	< 0.05
Chromaticism	-0.27	0.03	-8.09	< 0.01
Phrase Length	-0.53	0.28	-1.91	< 0.1
Overall model: $R = 0.92$, $R_{adj}^2 = 0.81$, $F(5, 22) = 25.04$, $p < 0.01$				

Table 1.3: Multiple regression results for the mean success ratings of each test melody.

More positive values of the predictors indicate greater deviation from the standards of the dataset (for pitch range and centre) or increased melodic complexity (for the remaining predictors), so we expect each predictor to show 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 1.3. The overall model accounts for approximately 85% of the variance in the mean success ratings. Apart from rhythmic structure, at least one predictor from each category made at least a marginally significant contribution to the fit of the model. Coefficients of all the selected predictors are negative as predicted. Overall, the model indicates that the judged success of a stimulus decreases as its pitch range and centre depart from the mean range and centre of the dataset, with increasing numbers of dissonant intervals and chromatic tones and if it has unequal phrase lengths.

1.4.6 Improving the Computational Systems

The constraints identified above mainly concern pitch range, intervallic structure and tonal structure. To examine whether the Systems can be improved to respect such constraints, we added several viewpoints 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 confusion of relative minor and major modes is due to the failure of any of the Systems to represent mode, so we added appropriate features to examine this hypothesis. We also hypothesise that the skewed distribution of pitch classes at phrase beginnings and endings can be better modelled by linked features representing scale degrees at phrase beginnings and endings. Finally, on the hypothesis that intervallic structure is constrained by tonal structure, we included a further feature representing an interaction between pitch interval and scale degree.

To examine whether the Systems can be improved to respect such constraints, we added the four selected features to the feature selection set used for System C. We ran the same feature selection algorithm over this extended feature space to select feature subsets which improve prediction performance; the results are given by Pearce and Wiggins (2007). In general, the resulting multiple-feature System, D, showed a great deal of overlap with System C: just three of the nine features present in System C were not selected for inclusion in System D. However, three of the four new features were selected for inclusion in System D. Ultimately, System D exhibits a lower average information content ($H = 1.91$) than System C ($H = 1.95$) in predicting unseen compositions in the dataset. The significance of this difference was confirmed by paired t tests over all 185 chorale melodies: $t(184) = 6.00$, $p <$ 0.01.

1.4.7 A Melody Generated by System D

We now present preliminary results on System D's capacity to generate

stylistically successful chorale melodies. We used it to generate several melodies, as described above, with the same base melodies.

System D: *Jesu, meiner Seelen Wonne*

Figure 1.3: Melody generated by System D, based on the same chorale as Figure 1.2.

Figure 1.3 shows System D's most successful melody, based on Chorale 365. Its tonal and melodic structure are much more coherent than System C's melodies. Our multiple regression model, developed above to account for the judges' ratings of stylistic success, predicts that this melody would receive a rating of 6.4 on a seven-point scale of success as a chorale melody. While this result is positive, other melodies were less successful; System D must be analysed using our method to examine its ability to *consistently* compose stylistically successful melodies.

1.4.8 Discussion and conclusions of the experiment

Our statistical finite context grammars did not meet the computational demands of chorale melody composition, regardless of the representational primitives used. Since we attempted to address the limitations of previous context-modelling approaches to generating music, we might conclude that more powerful grammars are needed for this task. However, other approaches are possible. Further analysis of the capacities of finite context modelling systems may prove fruitful: future research should use the methodology developed here to analyse System D, and identify and correct its weaknesses. The MCMC generation algorithm may be responsible for failure, rather than the limitation of the models to finite context representations of melodic structure: more structured generation strategies, such as pattern-based sampling techniques, may be able to conserve phrase-level regularity and repetition in ways that our Systems were not.

Our evaluation method also warrants discussion. The adapted CAT yielded insightful results for ratings of stylistic success even though the judges were encouraged to rate the stimuli according to an absolute standard (cf. Amabile,

1996). However, the results suggest possible improvements: first, avoid any possibility of method artefacts by randomising the presentation order of both test and practice items for each judge and also the order in which rating scales are presented; second, the judges' comments sometimes reflected aesthetic judgements, so they should also give ratings of aesthetic appeal, to delineate subjective dimensions of the product domain in the assessment (Amabile, 1996); and third, though influence of prior familiarity with the test items was ambiguous, bias resulting from recognition should be avoided.

Our results suggest that the task of composing a stylistically successful chorale melody presents significant challenges as a first step in modelling cognitive processes in composition. Nonetheless, our evaluation method proved fruitful in examining the generated melodies in the context of existing pieces in the style. It facilitated empirical examination of specific hypotheses about the models through detailed comparison of the generated and original melodies on several dimensions. It also permitted examination of objective features of the melodies which influenced the ratings and subsequent identification of weaknesses in the Systems and directions for improving them. This practically demonstrates the utility of analysis by synthesis for evaluating cognitive models of composition—if it is combined with an empirical methodology for evaluation such as that developed here.

1.5 Summary and Conclusions

In this chapter, our aim has been to connect music cognition and creativity conceptually and computationally by covering the issues that arise when we try to model human cognition and behaviour in these domains. We reviewed the literature on cognitive modelling and paid special attention to existing computational cognitive models of perception and composition. We also summarised an evaluation methodology (CAT) for systems that perform creative tasks. This methodology chimes well with the Creative Systems Framework (CSF) which distinguishes a set of rules, *R*, according to which a creative product can be constructed, a rule set, *E* which is used for evaluation of creative output, and a set of rules, *T* which can be used to traverse *R*.

The computational cognitive model that we have examined in some detail is based on an unsupervised machine learning paradigm and was originally constructed as a general model of human melodic learning. When applied to the prediction of melodic expectation, the basic model shows a performance superior to other models specifically designed to predict melodic expectation. When extended to the further task of melody segmentation, the resulting meta-model's performance is comparable with specialist segmentation models. These results point to the fact that the model seems to capture a more general, underlying perceptual mechanism and we suggest, therefore and because it is based on unsupervised machine learning, that it may be an explanatory model for both these cognitive phenomena, at this level of abstract representation.

Finally, we applied the pitch expectation model to the creative task of melody

composition and evaluated the generated melodies in a user study applying a variant of the CAT evaluation methodology. Despite its failure to consistently compose melodies indistinguishable from original melodies of the target style, the system has produced a number of acceptable melodies. We argue that the percentage of 'good melodies' among its output is of less importance than the fact any acceptable melodies were produced by an unsupervised learning model that lacks any pre-defined knowledge of musical structure but gains its knowledge exclusively through unsupervised learning on a training set. This means that we cannot be accused of showing the model how to appear to be creative and makes it applicable to any musical style or corpus.

We believe this to be the first time that an empirically validated computational cognitive model based on unsupervised machine learning has been used as the defining context (R) for a creative system, and it is therefore satisfactory that we have any good results at all, especially given that we have made no attempt to model T or E in a realistic way. This work is only a beginning, but, taken together, the results of the perceptual modelling and the melody production task suggest that the methods and approach presented here constitute a productive, general framework for the study of computational creativity. It provides clear directions for the future, which we expect to generate interesting empirical and theoretical developments in our ongoing research.

References and endnotes

Amabile, T. M. (1996). *Creativity in Context*. Westview Press, Boulder, Colorado.

Ames, C. and Domino, M. (1992). Cybernetic Composer: An overview. In Balaban, M., Ebciolu, K., and Laske, O., editors, *Understanding Music with AI: Perspectives on Music Cognition*, pages 186–205. MIT Press, Cambridge, MA.

Baroni, M. (1999). Musical grammar and the cognitive processes of composition. *Musicæ Scientiæ*, 3(1):3–19.

Baroni, M., Dalmonte, R., and Jacoboni, C. (1992). Theory and analysis of European melody. In Marsden, A. and Pople, A., editors, *Computer Representations and Models in Music*, pages 187–206. Academic Press, London.

Bharucha, J. J. (1987). Music cognition and perceptual facilitation: A connectionist framework. *Music Perception*, 5(1):1–30.

Bharucha, J. J. and 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. and Stoeckig, K. (1987). Priming of chords: Spreading activation or overlapping frequency spectra? *Perception and Psychophysics*, 41(6):519–524.

Boden, M. (1990). *The Creative Mind*. Abacus.

Boden, M. (1998). Creativity and artificial intelligence. *Journal of Artificial Intelligence*, 103(2):347–356.

Boltz, M. G. and Jones, M. R. (1986). Does rule recursion make melodies easier to reproduce? If not, what does? *Cognitive Psychology*, 18(4):389–431.

Bown, O. and Wiggins, G. A. (2008). From maladaptation to competition to cooperation in the evolution of musical behaviour. *Musicae Scientiae*. Special Issue on Evolution of Music. In press.

Bregman, A. S. (1990). *Auditory Scene Analysis*. The MIT Press, Cambridge, MA.

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34(1-3):71–105.

Cambouropoulos, E. (1996). A formal theory for the discovery of local boundaries in a melodic surface. In *Proceedings of the III Journées d'Informatique Musicale*, Caen, France.

Cambouropoulos, E. (2001). The local boundary detection model (LBDM) and its application in the study of expressive timing. In *Proceedings of the International Computer Music Conference*, pages 17–22, San Francisco. ICMA.

Cohen, P. R., Adams, N., and Heeringa, B. (2007). Voting experts: An unsupervised algorithm for segmenting sequences. *Intelligent Data Analysis*, 11(6):607–625.

Conklin, D. (2002). Representation and discovery of vertical patterns in music. In Anagnostopoulou, C., Ferrand, M., and Smaill, A., editors, *Proceedings of the Second International Conference of Music and Artificial Intelligence*, volume 2445 of *Lecture Notes in Computer Science*, pages 32–42. Springer, Berlin.

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*, pages 30–35, Brighton, UK. SSAISB.

Conklin, D. and Witten, I. H. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24:51–73.

Cope, D. (2005). *Computer Models of Musical Creativity*. Cambridge, MA: MIT Press.

Cross, I. (2007). Music and cognitive evolution. In Dunbar, R. and Barrett, L., editors, *Handbook of Evolutionary Psychology*, pages 649–667. Oxford University Press.

Cuddy, L. L. and Lunny, C. A. (1995). Expectancies generated by melodic intervals: Perceptual judgements of continuity. *Perception and Psychophysics*, 57:451–62.

Cutting, J. E., Bruno, N., Brady, N. P., and 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.

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.

Deliège, I. and Wiggins G. A. (2006). *Musical creativity: Current research in theory and practice*. Hove, UK: Psychology Press.

Deutsch, D. (1980). The processing of structured and unstructured tonal sequences. *Perception and Psychophysics*, 28(5):381–389.

Deutsch, D. and Feroe, J. (1981). The internal representation of pitch sequences in tonal music. *Psychological Review*, 88(6):503–522.

Ebciolu, K. (1988). An expert system for harmonizing four-part chorales. *Computer Music Journal*, 12(3):43–51.

Eerola, T. and North, A. C. (2000). Expectancy-based model of melodic complexity. In Woods, C., Luck, G., Brochard, R., Seddon, F., and Sloboda, J. A., editors, *Proceedings of the Sixth International Conference on Music Perception and Cognition*, Keele, UK. Keele University.

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

Hall, M. and Smith, L. (1996). A computer model of blues music and its evaluation. *Journal of the Acoustical Society of America*, 100(2):1163–1167.

Hiller, L. and Isaacson, L. (1959). *Experimental Music*. McGrawÐHill, New York.

Honing, H. (2007). Preferring the best fitting, least flexible, and most surprising prediction: Towards a bayesian approach to model selection in music cognition. In *Proceedings of the Society for Music Perception and Cognition (SMPC)*, Concordia University, Montreal, Canada.

Huron, D. (2001). Tone and voice: A derivation of the rules of voice-leading from perceptual principles. *Music Perception*, 19(1):1–64.

Huron, D. (2006). *Sweet Anticipation: Music and the Psychology of Expectation*. Bradford Books. MIT Press, Cambridge, MA.

Jackendoff, R. (1987). *Consciousness and the Computational Mind*. MIT Press, Cambridge, MA.

Johnson-Laird, P. N. (1991). Jazz improvisation: A theory at the computational level. In Howell, P., West, R., and Cross, I., editors, *Representing Musical Structure*, pages 291–325. Academic Press, London.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, New Jersey.

Justus, T. and Hutsler, J. J. (2005). Fundamental issues in the evolutionary psychology of music: Assessing innateness and domain specificity. *Music Perception*, 23(1):1–27.

Koestler, A. (1964). *The Act of Creation*. Hutchinson & Co., London.

Krumhansl, C. L. (1990). *Cognitive Foundations of Musical Pitch*. Oxford University Press, Oxford.

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., Toivanen, P., Eerola, T., Toiviainen, P., Järvinen, T., and Louhivuori, J. (2000). Cross-cultural music cognition: Cognitive methodology applied to North Sami yoiks. *Cognition*, 76(1):13–58.

Large, E. W., Palmer, C., and Pollack, J. B. (1995). Reduced memory representations for music. *Cognitive Science*, 19(1):53–96.

Lerdahl, F. (1988). Cognitive constraints on compositional systems. In Sloboda, J. A., editor, *Generative Processes in Music: The Psychology of Performance, Improvisation and Composition*, pages 231–259. Clarendon Press, Oxford.

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

MacKay, D. J. C. (1998). Introduction to Monte Carlo methods. In Jordan, M. I., editor, *Learning in Graphical Models*, NATO Science Series, pages 175–204. Kluwer Academic Press, Dordrecht, The Netherlands.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Manzara, L. C., Witten, I. H., and James, M. (1992). On the entropy of music: An experiment with Bach chorale melodies. *Leonardo*, 2(1):81–88.

Marr, D. (1982). *Vision*. W. H. Freeman, San Francisco.

Marsden, A. (2000). *Representing Musical Time: A Temporal-Logic Approach*. Swets & Zeitlinger, Lisse.

McClamrock, R. (1991). Marr's three levels: A re–evaluation. *Minds and Machines*, 1:185–196.

Meredith, D. (2006). The ps13 pitch spelling algorithm. *Journal of New Music Research*, 35(2):121–159.

Meyer, L. B. (1957). Meaning in music and information theory. *Journal of Aesthetics and Art Criticism*, 15(4):412–424.

Mithen, S. J. (2006). *The Singing Neanderthals: The Origins of Music, Language, Mind, and Body*. Harvard University Press, Cambridge, MA.

Müllensiefen, D., Pearce, M. T., and Wiggins, G. A. (2008). Melodic segmentation: A new method and a framework for model comparison. In *Proceedings of ISMIR 2008*. In review.

Narmour, E. (1990). *The Analysis and Cognition of Basic Melodic Structures: The Implication-Realistation Model*. The University of Chicago Press, Chicago.

Narmour, E. (1992). *The Analysis and Cognition of Melodic Complexity*. The University of Chicago Press.

Palmer, C. and Krumhansl, C. L. (1990). Mental representations for musical metre. *Journal of Experimental Psychology: Human Perception and Performance*, 16(4):728–741.

Pearce, M. T. (2005). *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. PhD thesis, Department of Computing, City University, London, UK.

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

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

Pearce, M. T., Müllensiefen, D., Lewis, D., and Rhodes, C. S. (2007). David Temperley, *Music and Probability*. Cambridge, Massachusetts: MIT Press, 2007, ISBN-13: 978-0-262-20166-7 (hardcover) $40.00. *Empirical Musicology Review*, 2(4):155–163.

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

Pearce, M. T. and Wiggins, G. A. (2006). Expectation in melody: The influence of context and learning. *Music Perception*, 23(5):377–406.

Pearce, M. T. and Wiggins, G. A. (2007). Evaluating cognitive models of musical composition. In Cardoso, A. and Wiggins, G. A., editors, *Proceedings of the 4th International Joint Workshop on Computational Creativity*.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, 46(1):77–105.

Ponsford, D., Wiggins, G. A., and Mellish, C. S. (1999). Statistical learning of harmonic movement. *Journal of New Music Research*, 28(2):150–77.

Potter, K., Wiggins, G. A., and Pearce, M. T. (2007). Towards greater objectivity in music theory: Information-dynamic analysis of minimalist music. *Musicae Scientiae*, 11(2):295–324.

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285.

Ritchie, G. (2007). Some empirical criteria for attributing creativity to a computer program. *Minds and Machines*, 17(1):67–99.

Robertson, J., de Quincey, A., Stapleford, T., and Wiggins, G. A. (1998). Real-time music generation for a virtual environment. In Nack, F., editor, *Proceedings of the ECAI'98 Workshop on AI/Alife and Entertainment*, Brighton, England.

Rutherford, J. and Wiggins, G. A. (2002). An experiment in the automatic creation of music which has specific emotional content. In *Proceedings of the 2002 International Conference on Music Perception and Cognition*. AMPS

and Causal Productions.

Saffran, J. R. (2003). Absolute pitch in infancy and adulthood: The role of tonal structure. *Developmental Science*, 6(1):37–49.

Saffran, J. R. and Griepentrog, G. J. (2001). Absolute pitch in infant auditory learning: Evidence for developmental reorganization. *Developmental Psychology*, 37(1):74–85.

Saffran, J. R., Johnson, E. K., Aslin, R. N., and Newport, E. L. (1990). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70:27–52.

Schaffrath, H. (1995). The Essen folksong collection. In Huron, D., editor, *Database containing 6,255 folksong transcriptions in the Kern format and a 34-page research guide [computer database]*. CCARH, Menlo Park, CA.

Schellenberg, E. G. (1996). Expectancy in melody: Tests of the implication-realisation model. *Cognition*, 58(1):75Ð125.

Schellenberg, E. G. (1997). Simplifying the implication-realisation model of melodic expectancy. *Music Perception*, 14(3):295–318.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423 and 623–656.

Simon, H. A. and Sumner, R. K. (1968). Pattern in music. In Kleinmuntz, B., editor, *Formal Representation of Human Judgement*, pages 219–250. Wiley, New York.

Sloboda, J. (1985). *The Musical Mind: The Cognitive Psychology of Music*. Oxford Science Press, Oxford.

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*. MIT Press, Cambridge, MA.

Temperley, D. (2003). Communicative pressure and the evolution of musical styles. *Music Perception*, 21(3):313–337.

Temperley, D. (2007). *Music and Probability*. MIT Press, Cambridge, MA.

Thompson, W. F., Cuddy, L. L., and 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.

Triviño-Rodriguez, J. L. and Morales-Bueno, R. (2001). Using multi-attribute prediction suffix graphs to predict and generate music. *Computer Music Journal*, 25(3):62–79.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York.

von Hippel, P. T. (2000). Redefining pitch proximity: Tessitura and mobility as constraints on melodic intervals. *Music Perception*, 17(3):315–127.

Wallas, G. (1926). *The Art of Thought*. Harcourt Brace, New York.

Wiggins, G. A. (2006a). A preliminary framework for description, analysis and comparison of creative systems. *Journal of Knowledge Based Systems*, 19(7):449–458.

Wiggins, G. A. (2006b). Searching for computational creativity. *New Generation Computing*, 24(3):209–222.

Wiggins, G. A. (2007). Models of musical similarity. *Musicae Scientiae, Discussion Forum 4a*.

Wiggins, G. A., Miranda, E., Smaill, A., and Harris, M. (1993). A framework for the evaluation of music representation systems. *Computer Music Journal*, 17(3):31–42. Machine Tongues series, number XVII; Also from Edinburgh as DAI Research Paper No. 658.

---

[1] And we do mean "relatively" – it is absolutely clear that this is an over-simplification. However, one has to start somewhere.

[2] Often therefore, we are able to learn more about brain operation from pathological cases (e.g., brain-damaged patients) than from normal ones.

[3] Improvisation may be seen as a special case of composition where the composer is the performer and is subject to extra constraints of immediacy and fluency (Sloboda, 1985).

[4] We do not propose Metropolis sampling as a cognitive model of melodic comparison, but use it merely as a means of generating melodies which reflect the internal state of knowledge and capacities of the trained model.