8 Computational Models of Music Cognition

David Temperley

Eastman School of Music, Rochester, New York

I. Introduction

In recent decades, computational research has assumed an increasingly important role in the study of cognition. Computer modeling is generally regarded as one of the three main approaches—along with experimental psychology and neuroscience—that comprise the interdisciplinary field of "cognitive science." It is no surprise, then, that computational work has become an important part of the field of music cognition as well. In this chapter, I present an overview of this research. I discuss some of the main problems that have been addressed, some approaches that have been taken in addressing them, and some general issues that arise in the computational modeling of music cognition.

Probably most readers have an intuitive sense of what is meant by "computational model." The exact meaning of this phrase is, however, rather subtle and deserves some scrutiny. Let us begin with the word "model." A model is a description of something; a cognitive model is a description of some aspect of cognition (usually, human cognition). A computational system is only a cognitive model if it is intended, or used, in this way. This is a crucial point to bear in mind. In many cases—in music as in other domains—computer systems are devised to perform tasks that are also performed by humans, but without any intent of modeling the human cognitive process; the intent is simply to get the job done—to perform the task as effectively and efficiently as possible. (Consider a calculator, for example.) Generally speaking, we might describe such work as computational *engineering* rather than cognitive science. The cognitive processes discussed in this chapter—such as key identification, meter identification, and composition—have sometimes been approached from a purely engineering point of view, not with the aim of modeling cognition but simply to serve some practical purpose.

Having said all this, it is generally agreed that simply "getting the job done" can be a useful approach to cognitive modeling as well. That is to say: In trying to understand how the human cognitive system solves a particular problem, it is often useful to begin by asking what needs to be done to solve the problem from a purely computational point of view. Indeed, this was the original rationale for the

The Psychology of Music. DOI: http://dx.doi.org/10.1016/B978-0-12-381460-9.00008-0

computational approach to cognition—a rationale articulated most famously, perhaps, by Marr (1982). In some cases, this approach has led to important insights into cognitive processes. In this sense, there is an overlap—a convergence, one might say—between the cognitive and engineering sides of computational research; ideas proposed on one side may sometimes be useful on the other. Still, the ultimate goals of the two sides are clearly distinct. Although the ability of a cognitive model to "get the job done" may be one criterion in evaluating it, other considerations are also relevant—in particular, experimental or neurological evidence that bears on the cognitive plausibility of the system. We will sometimes consider such evidence in evaluating the models discussed here.

A further point is needed about the term "model." Many studies in music cognition propose a relationship between properties of a stimulus and some aspect of musical behavior or experience, often using techniques of regression analysis. Although some might consider such proposals to be models, I will generally not do so here. The term "cognitive model" usually implies not only a relationship between input and output but also a claim about the cognitive process whereby that output is produced. It is at least doubtful that a regression analysis, on its own, implies any such claim. In addition, as a practical matter, regarding such studies as models would require us to consider a large proportion of the research in music cognition, far more than could properly be surveyed in a single chapter. The line must be drawn somewhere!

The word "computational" is also more subtle than it might first appear. At first thought, one might assume that a computational model is simply a model that is implemented on a computer. But this proves not to be a very useful criterion for categorizing models. A few of the models presented here are so simple that they may be quite easily implemented with pencil and paper, without need for a computer; such simplicity should surely be considered a virtue rather than a disqualification. To call a model "computational" implies, rather, that it is specified in a precise, complete, and rigorous way—such that it *could* be implemented on a computer. Computer implementation is useful, in part, because it ensures that this requirement is met. In my own experience, the process of implementing a model has often drawn my attention to aspects of it that were underspecified or inconsistent. Computer implementation has other benefits as well, making the development and testing of models much easier, faster, and less prone to error; the complexity of many models proposed in recent years makes the help of computers almost indispensable. (See also Oxenham, Chapter 1, on models of pitch perception, and Honing, Chapter 9, on models of timing.)

Most computational cognitive models describe cognitive processes at a fairly abstract—some might say "computational"—level, without regard for the way these processes are implemented in the neurological hardware of the brain. It is now well established in cognitive science that a computational description of a cognitive process is no less "real" than a neurological one; it simply represents a more abstract level of description. In recent years, computational models have also been proposed for neurological processes, but this approach has not been widely applied to music.

What follows is a survey of some important research in computational modeling of music cognition. We begin with problems of perception or information processing—problems of extracting various kinds of information from music as it is heard. Here we focus primarily on two especially well-studied problems, key-finding and meter-finding, but briefly consider several other problems as well. We then turn to three other broad issues: the modeling of musical experience, the modeling of performance, and the modeling of composition.

II. Models of Key-Finding

Key plays a central role in the understanding of Western tonal music. The key of a piece provides the framework in which the functions of individual pitches are understood; for example, to identify a pitch as the tonic or the leading tone presupposes that the key has been identified. Experimental research has suggested that listeners in general—even those without extensive musical training—are sensitive to key; for example, given a tonal context, listeners generally judge notes within the scale of the context key to "fit" better than those that are not (Cuddy, 1997; Krumhansl, 1990). In light of the importance of key and its well-established psychological reality, the question of how listeners identify key—sometimes known as the "key-finding" problem—is of great interest, and it is not surprising that it has attracted attention in computational modeling research.

With any musical information-processing model, it is important to consider the kind of input representation that is assumed. Especially important is the distinction between signal-level (or "audio") representations, which take direct sound input as produced in a performance, and symbolic representations, which require that some symbolic information be extracted from the input before processing begins. To date, nearly all computational models in music cognition—in key-finding as in other areas—have assumed symbolic input. (By contrast, many systems designed for practical use—such as systems designed for categorizing or identifying music on the internet—assume audio input.) In many cases, a piece is represented simply as an array of notes with pitches and time points—what is sometimes called a "piano-roll" or "MIDI" representation. Arguments can be made for both signal-level and symbolic approaches. At first thought, the signal-level approach might seem more defensible; clearly, this more accurately represents what the listener encounters, at least initially. Arguments for the symbolic approach might be made as well, however. Many information-processing problems can be solved much more easily and effectively from symbolic information. (In key identification, for example, it is the notes that matter; signal-level properties such as timbre and dynamics are mostly irrelevant.) There is also evidence that listeners do form a symbolic representation of some kind (most listeners can extract note information from a piece to some extent—for example, singing back the melody); therefore it seems reasonable to assume such a representation as a starting point in cognitive modeling.

Our survey of key-finding models begins with the classic work of Longuet-Higgins and Steedman (1971). Longuet-Higgins and Steedman propose a key-finding model based on the conventional association between keys and scales. In the Longuet-Higgins/Steedman (hereafter LH-S) model, a melody is processed one note at a time in a "left-to-right" fashion (only monophonic input is allowed). At each note, all the keys whose scales do not contain that note are eliminated from consideration. (For minor keys, the harmonic minor scale is assumed, though notes of the melodic minor are allowed if used in an appropriate context.) If, at any point, only one key remains, that is the chosen key. If, at any point, all keys have been eliminated, the model undoes the previous step; then, from among the remaining eligible keys, it chooses the one whose tonic pitch is the first note of the piece (or, failing that, the key whose dominant pitch is the first note). If at the end of the melody there is more than one eligible key remaining, the "first-note" rule again applies. Longuet-Higgins and Steedman tested their model on the 48 fugue subjects of Bach's *Well-Tempered Clavier*; the model obtained the correct result in all 48 cases.

Despite its good performance on the *Well-Tempered Clavier*, it is not difficult to find cases where the LH-S model does not perform so well. The melodies in Figures 1a and 1b, both in C major, illustrate some problems with the model and also represent some general challenges posed by the key-finding problem. In Figure 1a, the model would reach the end of the phrase with two eligible keys remaining, C major and G major, and would incorrectly choose G major because the melody begins with G. In Figure 1b the first F♯ would cause the correct key, C major, to be eliminated, leaving E minor as the only eligible key at the end of the

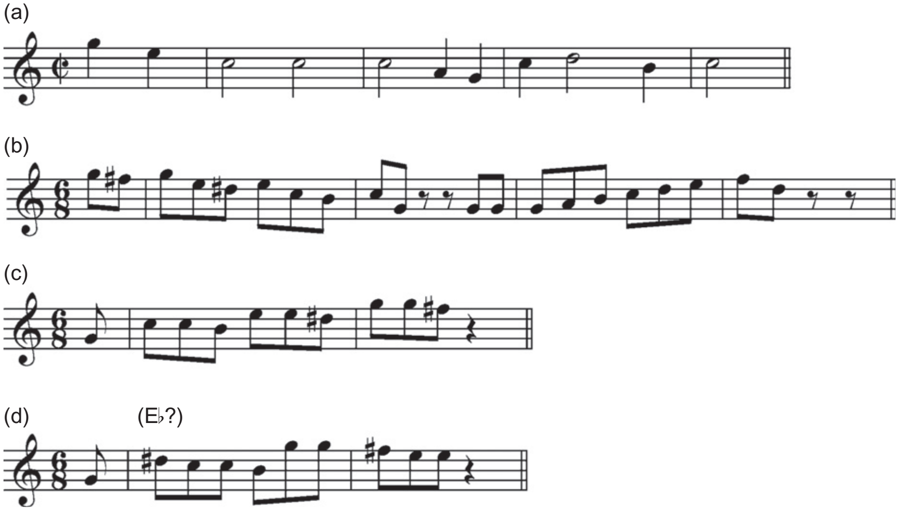

Figure 1 (a) George M. Cohan, "You're a Grand Old Flag." (b) "Mexican Hat Dance." (c) The pitches of the first phrase of Figure 1B, rearranged in a different order. (d) Another rearrangement of the pitches in the first phrase of Figure 1b.

first phrase. In general, the model has difficulty in cases where a melody does not use a complete scale or uses notes outside the scale.

The key-finding model of Krumhansl and Schmuckler, presented most fully in Krumhansl (1990), builds on the LH-S model but also addresses some of its shortcomings. The Krumhansl-Schmuckler (K-S) model is based on the concept of a "key profile," a vector of 12 values representing the stability or appropriateness of each pitch class in relation to a key. The key profiles were derived from experiments (Krumhansl & Kessler, 1982) in which listeners had to rate how well each pitch class "fit" with a tonal context (Figure 2). The profiles represent well-established principles of music theory, with notes of the tonic triad rated most highly, followed by other notes of the scale

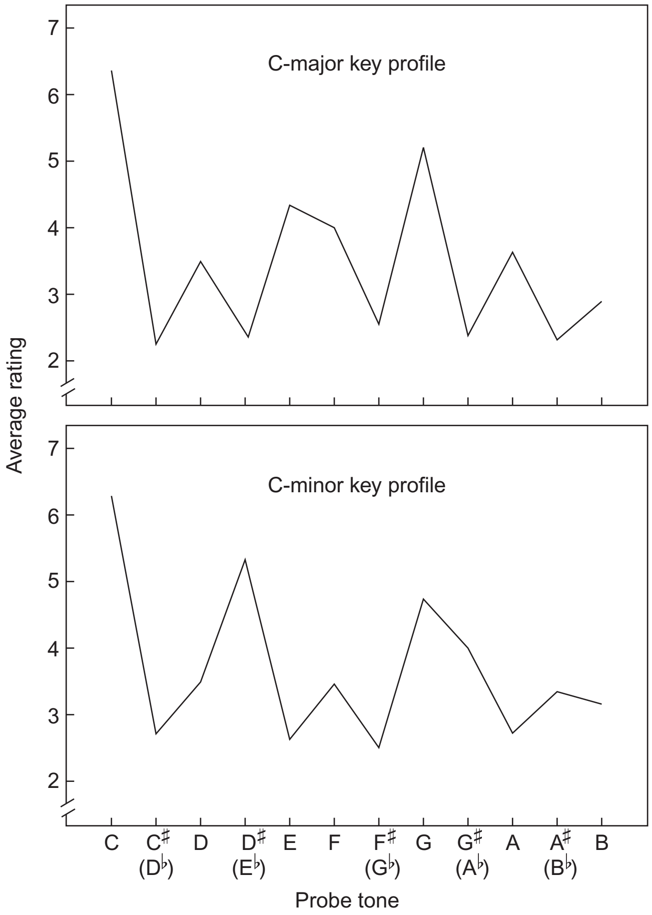

Figure 2 Ratings for individual tones ("probe tones") in a context of C major (top) and C minor (bottom).
From Krumhansl and Kessler (1982). ©1982 American Psychological Association.

(major or natural minor), followed by chromatic pitches. Given these key profiles, the model then generates an "input vector" for the piece; this is again a 12-valued vector showing the total duration of each pitch class in the piece. The model finds the correlation between each key profile and the input vector, and chooses the key whose profile yields the highest correlation. As Krumhansl observes, this is a kind of "template matching": If the most frequent pitches in the piece are highly rated in the key profile, the corresponding correlation value will be high. It can be seen that such a model could, in principle at least, address the problems illustrated by Figure 1. In Figure 1a, although all the notes of the melody are contained within the C-major and G-major scales, the fact that the notes are mostly tonic-triad notes in C major but not in G major will give an advantage to the former key. In Figure 1b, though some of the notes are not contained in the C-major scale, the preponderance of C-major tonic-triad notes again favors that key over other possibilities.

The K-S model might be described as a "distributional" model of key-finding, because it judges key on the basis of the distribution of pitch classes in a piece, without regard for their temporal arrangement or register. Much of the subsequent work on key-finding has followed this distributional approach. In the model of Vos and Van Geenen (1996), each key receives points for pitches that belong to its scale, tonic triad, or dominant seventh chord. In the model of Chew (2002), pitches are arranged in a spiral representational space; each key has a characteristic point in the space, and the preferred key is the one closest to the mean position of all the pitches. Temperley (2007) proposes a probabilistic construal of the K-S model, in which a key profile represents the expected distribution of pitch classes given a key (this probabilistic approach will be discussed further later). Leman (1995) presents a key-finding model for audio input; the model employs 12-valued vectors very similar to the key profiles of the K-S model. Vectors are constructed for short segments of the input, representing the salience of each pitch class; these are then correlated with vectors representing the characteristic pitch-class content of each key.

A major challenge for distributional key-finding models is how to handle modulations, or changes of key within a piece. Shmulevich and Yli-Harja (2000) propose a variant of the K-S model in which the input is divided into small time slices, and slices with similar input vectors are then grouped together to form larger key sections. In the model of Huron and Parncutt (1993), the input vector is recalculated from moment to moment with previous events weighted under an exponential decay; thus shifts in the pitch-class content of the piece may result in changes of key. Toiviainen and Krumhansl (2003) offer yet another approach; in this study, a neural network is trained to simulate a spatial representations of keys, and changes in pitch-class content are reflected in the activation pattern of the network.

The distributional approach to key-finding is remarkably powerful: recent distributional models have achieved high accuracy rates on musical corpora (using materials such as classical pieces and folk melodies). With respect to human key-finding, however, this view is clearly somewhat oversimplified. Consider Figure 1c; this shows the pitches of the first phrase of Figure 1b arranged in a different order. While the original phrase clearly projects a key of C major, the reordered version seems to imply E minor. Experiments have confirmed that the same pitches in different orders can have different key implications (Matsunage & Abe, 2005). The question then arises, what other kinds of information are used in human key detection? A number of piecemeal answers to this question have been proposed. It has been suggested that certain conventional pitch patterns are especially important for key-finding: a rising fourth suggests scale degrees $\hat{5}$ to $\hat{1}$; a descending tritone suggests scale degrees $\hat{4}$ to $\hat{7}$(Butler, 1989; Vos, 1999). (See Deutsch, Chapter 7, for further discussion.) To date, however, few concrete proposals have been offered for how such "structural" cues might be incorporated into a testable key-finding model. The solution may also lie partly in harmony; one difference between Figures 1b and Figure 1c is that they imply different harmonic progressions (C major in the first case, E minor to B major in the second). Of interest in this connection are recent models that analyze key and harmony simultaneously; these will be discussed in Section IV.

III. Models of Meter-Finding

Like key perception, the perception of meter is a crucial part of musical understanding. Meter impacts a variety of other aspects of musical experience: for example, it affects the perceived complexity of melodic patterns (Povel & Essens, 1985), the perceived similarity between patterns (Gabrielsson, 1973), the stability of events (Palmer & Krumhansl, 1987), temporal expectancy (Jones, Moynihan, MacKenzie, & Puente, 2002), and performance errors (Palmer & Pfordresher, 2003). Studies of these effects have also provided ample evidence for the psychological reality of meter. Thus meter-finding is a natural problem to address from a computational viewpoint.

What exactly *is* meter? Although there is some diversity of opinion as to how meter should be characterized, there is general agreement on certain essential points. Fundamentally, meter consists of beats or pulses—points in time, subjectively accented in the mind of the listener and not necessarily always coinciding with events in the music (though they are *inferred* from events in the music). Beats vary in subjective accentuation or "strength," forming a multileveled hierarchical structure. An intermediate level of the hierarchy represents the main beat or "tactus," the level of beats at which one naturally taps or conducts. Some metrical models produce only a single level of beats (generally the tactus level), while others produce additional lower and higher levels. Metrical structure is conveyed by standard music notation (up to the level of the measure anyway) and can be represented in this way; it can also be represented, more explicitly, in what is known as a "metrical grid." Figure 3 shows the two melodies in Figures 1a and 1b along with their metrical grids.

Longuet-Higgins and Steedman's 1971 study—discussed earlier with regard to key-finding—also proposes a meter-finding model, and once again, it provides a good starting point for our discussion. Like their key-finding model, the LH-S

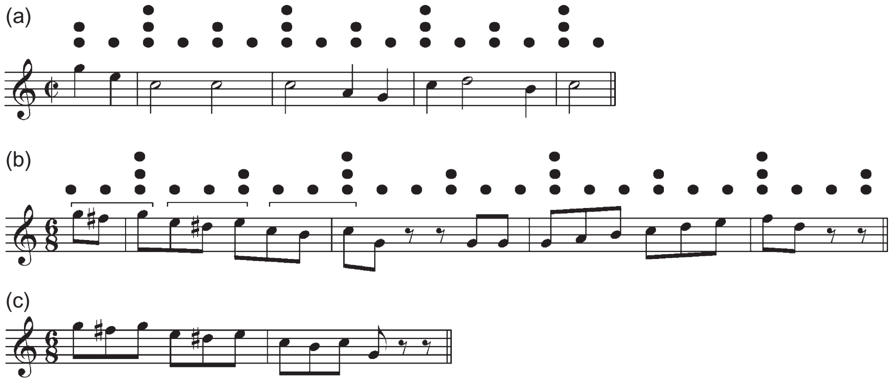

Figure 3 (a) The melody in Figure 1a, with metrical grid. (b) The melody in Figure 1b, with metrical grid. (c) The first phrase of Figure 3b, showing an alternative meter.

meter-finding model assumes monophonic input and processes a melody one note at a time. The input is simply "the relative durations of the notes and rests, as they would be given in the score" (p. 226). The metrical structure is constructed in a "bottom-up" fashion: the first note of the piece is assumed to define a metrical level (with beats at its beginning and end points), and then each new, longer note defines a new higher level. In Figure 3a, for example, the first note defines a quarter-note level and the third note defines a half-note level. As with key-finding, the LH-S meter model assumes a principle of "congruence," which states that no event will conflict with the correct structure until that structure has been established. Figure 3a respects this principle: the syncopation in measure 3 conflicts with the notated meter (in that it suggests a half-note beat starting on the second quarter of the measure), but by this point, the correct meter has already been completely identified. Additional rules construct metrical levels when certain conventional rhythmic patterns are encountered; for example, a dactyl (a long note followed by two short notes) triggers a metrical level whose unit length is the entire dactyl.

The model is tested on Bach's 48 fugue subjects; its performance is impressive but far from perfect. Perhaps the greatest flaw of the model is that it has no mechanism for generating metrical levels below the level of the first note. The model is also challenged by melodies that are purely isochronous (i.e., in which all notes are the same duration); in such cases, the model defines the first duration as a metrical level but is not able to infer any higher levels. Subsequent studies have proposed refinements to the LH-S model, addressing these problems. The model of Lee (1991) allows the subdivision of previously established levels; Steedman (1977) proposes a way of incorporating parallelism—the preference for metrical levels that align with repeated melodic patterns.

A very different approach to meter-finding is represented by the model of Povel and Essens (1985). These authors propose that meter-finding consists

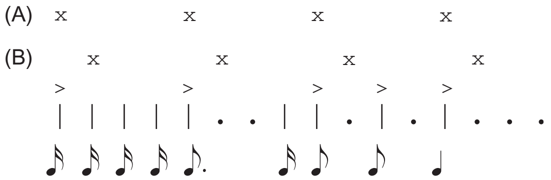

Figure 4 A rhythmic pattern (below). Accent marks show accented events according to Povel and Essens's (1985) model. The x's above the pattern show a compatible clock (a) and an incompatible one (b).

of considering all possible "clocks," where a clock is a level of beats superimposed on the input, and choosing the best one. (Unlike the LH-S model, this model only infers a single level of beats.) Like the LH-S model, the Povel-Essens (P-E) model assumes input that is essentially note durations as indicated in a score: each note is encoded as a multiple of a small rhythmic unit. For example, the rhythmic pattern shown in Figure 4 could be encoded as 1-1-1-1-3-1-2-2-4 (assuming sixteenth notes as the basic unit). (Strictly speaking, the values in the input refer not to literal durations but to the time intervals between the onset of one note and the onset of the next; this is more properly called inter-onset interval, or IOI.) A clock is then evaluated by the alignment of its beats with events and accented events; an accented event is defined as follows (accented events are underlined):

1. a note at the beginning or end of a sequence of 3 or more short notes (e.g., 2 $\underline{1}$ 1 1 . . . or 1 1 $\underline{2}$),
2. the second note in a cluster of two (e.g. 2 1 $\underline{2}$), or
3. a note with no note on either adjacent beat (e.g. 2 $\underline{2}$ 1).

More precisely, the model considers how many of the clock's beats *fail* to align with events and accented events, assigning a penalty for each beat that falls on an unaccented event and a higher penalty for beats that coincide with no event. The lower the total penalty, the more strongly the clock fits with, or implies, the pattern. The degree of fit between a clock and a pattern is called the "induction strength" of the clock. In Figure 4, clock (a) has a higher induction strength than clock (b) because all of its beats coincide with accented events. Povel and Essens also take the induction strength of a pattern's most strongly induced clock to be the induction strength of the pattern itself; this is taken as a predictor of the complexity of the pattern, with patterns of lower induction strength being more complex. (Povel and Essens also propose an alternative way of predicting the complexity of a pattern, based on how efficiently it can be encoded given a clock; this aspect of their study has been less influential and will not be considered further here.)

Povel and Essens tested their model experimentally. In one experiment, subjects heard rhythmic patterns of different levels of induction strength and had to reproduce them; patterns with lower induction strength were reproduced less accurately. In another experiment, subjects heard rhythmic patterns accompanied by a clock (a regular pulse) and had to indicate the simplicity of each pattern; the assumption was that subjects would judge a pattern to be simpler when it was accompanied by a compatible clock (i.e., one that was strongly implied by the pattern). The experiment showed a positive relationship between subjects' judgments of simplicity and the compatibility of the clock with the pattern.

The LH-S and P-E models illustrate a fundamental distinction between two approaches to computational modeling—a distinction that we will encounter several times in this chapter. The P-E model operates by considering a set of complete analyses of the entire input and evaluating them. The criteria for evaluating analyses may be called an *evaluation function* (borrowing a term from computer science); the model operates by finding the analysis that best satisfies the evaluation function. Such models may be called *optimization models*. The LH-S model, by contrast, processes a piece in a "left-to-right" manner, processing one note at a time. We cannot characterize the model in relation to an evaluation function (because none is involved); we can only say that it follows a procedure that leads to a certain analysis. We might call models of this kind *procedural models*. (A similar distinction may be made with regard to key-finding models: the LH-S key-finding model is a procedural model, whereas the K-S model is an optimization model.) In general terms, both optimization models and procedural models have points in their favor. Procedural models more clearly simulate the process of musical listening: In inferring the meter of a melody, we do not wait until we have heard the entire melody before making a judgment but rather begin forming our analysis immediately. On the other hand, optimization models might, in principle, account for this incremental listening process by applying the evaluation function to increasingly large portions of the piece as it unfolds in time. An advantage of optimization models is that the evaluation function in itself provides a succinct, yet precise, description of what the model is doing: It is finding the analysis that best satisfies the evaluation function. The question of how the model finds this best analysis (often a nontrivial question) is then a question of *search*, which might be solved in various ways or simply left unspecified.

Both the P-E and LH-S models are limited in one important way: they assume input in which the note durations are exact multiples of a small rhythmic unit (this is sometimes referred to as *quantized* input). This greatly simplifies the meter-finding problem, since it means that metrical structures, too, may be assumed to be perfectly regular; once a rhythmic value (such as the measure or quarter note) is determined, it can simply be projected through the rest of the piece. In fact, of course, human rhythmic performance is *not* quantized; it always deviates from perfect regularity, often intentionally. In early work on meter-finding, the problem of quantization—adjusting the continuously varying durations of real performance to be multiples of a common unit—was generally treated separately from the problem of meter-finding itself (Desain & Honing, 1989). More recently, a number of models have attempted to address quantization and meter-finding within a single framework.

One of the first models to address both quantization and meter-finding was that of Rosenthal (1992). Rosenthal's model finds pairs of note onsets in a piece (not necessarily adjacent) and takes the time interval between them to imply a metrical level; this is then extrapolated to the next beat, but there is some flexibility in the placement of the beat if the onset intervals are not completely regular. (The model can also assert a beat where no event occurs.) Many rhythmic levels are generated in this way and then grouped into "families" of levels that are related by simple ratios. The model of Dixon (2001, 2007) extends Rosenthal's approach. In Dixon's model, IOIs are "clustered" into potential rhythmic levels, as in Rosenthal's model, and then evaluated on the basis of their alignment with "salient" events; salience reflects the duration, amplitude, and pitch of notes, with higher-pitched events given greater salience.

Several recent models could be seen as building on the optimization approach of the Povel-Essens model, in that they operate by evaluating many possible analyses of the entire input. Parncutt (1994) proposes a simple model that takes into account listeners' strong preference for tactus levels in a certain absolute time range, centered on about 600 ms. McAuley and Semple (1999) note that the P-E model's "negative-evidence" approach—counting the number of beats in the clock that do not coincide with events—tends to favor clocks with fewer beats, while a positive-evidence model favors clocks with more beats; they also consider a "hybrid" model that combines positive and negative evidence. The model providing the best fit to experimental tapping data depends in a complex way on the level of experience of the subjects and the tempo of the pattern (see also Eck, 2001). In the model of Temperley and Sleator (1999), the evaluation function considers not only the alignment of beats with events, but the regularity of the metrical structure itself (the difference between each beat interval and the previous one); thus the model can accommodate unquantized input.

A recent trend in optimization models has been the application of Bayesian probabilistic techniques. From a probabilistic perspective, meter-finding can be construed as the problem of finding the structure maximizing $P(\text{structure} \mid \text{surface})$, where the structure is a metrical structure and the surface is a pattern of notes (or, for that matter, an audio signal). By Bayesian reasoning,

$$P(\text{structure} \mid \text{surface}) \propto P(\text{surface} \mid \text{structure}) \times P(\text{structure}) \qquad \text{(Eq. 1)}$$

Or, in the case of meter-finding:

$$P(\text{meter} \mid \text{note pattern}) \propto P(\text{note pattern} \mid \text{meter}) \times P(\text{meter}) \qquad \text{(Eq. 2)}$$

Thus the quantity of interest depends on the probability of the note pattern given the meter (known as the "likelihood" of the note pattern) and the "prior" probability of the meter. The prior probability of the meter might depend on things such as the absolute interval between beats (bearing in mind that tactus intervals near a certain ideal value are preferred), the relationship between beat levels (duple or triple), and the regularity of beats; the likelihood of the note pattern depends on how well the notes are aligned with the meter. As an illustration, we might apply this approach to the model of Povel and Essens. In the P-E model, all clocks are equal in prior probability (though one could also modify this assumption—for example, assigning higher probability to duple over triple clocks); in that case, $P(\text{clock} \mid \text{pattern}) \propto P(\text{pattern} \mid \text{clock})$, and the best clock is simply the one that generates the pattern with highest probability. To calculate $P(\text{pattern} \mid \text{clock})$, we could stipulate probabilities for an event occurring on a beat (say, .8; the probability of no note occurring on a beat is then $1-.8=.2$) and on a nonbeat (say, .4, or $1-.4=.6$ for no note on a nonbeat); the total probability of a pattern would be the product of these probabilities. This would correctly predict that clock (a) in Figure 4 is preferred over clock (b), as shown in Figure 5. A probabilistic model of this kind is *generative*, in that it calculates the probability of a structure given a surface by considering the probability of generating the surface from the structure.

The probabilistic approach to meter-finding has been applied in a variety of studies. Cemgil and colleagues (Cemgil, Desain, & Kappen, 2000; Cemgil, Kappen, Desain, & Honing, 2000) propose two probabilistic models for different aspects of the meter-finding process; one derives a single level of pulses from unquantized input, and another infers a multilevel metrical grid from quantized input. The model of Raphael (2002), which also takes quantized input, assumes a generative process in which a metrical position is chosen for each note dependent on the metrical position of the previous note; in this way, the model is able to capture the kinds of contextual accent proposed by Povel and Essens (for example, favoring long IOIs after notes on strong beats). The model of Temperley (2007) builds on these approaches, generating a full (three-level) metrical grid from unquantized input. Finally, Sadakata, Desain, and Honing (2006) propose a probabilistic model that engages more directly with experimental data. In an experiment, subjects heard a time interval divided into two IOIs in different ways and had to indicate the rhythmic notation they perceived for it; the model of Sadakata et al. predicts this data in a Bayesian fashion from corpus data (indicating the probability of different notated rhythmic patterns) and performance data (indicating the distribution of performed rhythms given a notated rhythm).

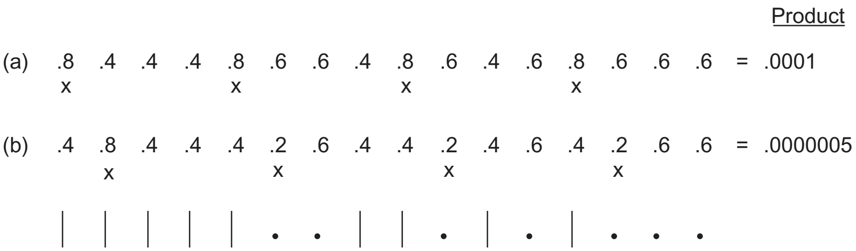
Product
(a) .8 .4 .4 .4 .8 .6 .6 .4 .8 .6 .4 .6 .8 .6 .6 .6 = .0001
x x x x
(b) .4 .8 .4 .4 .4 .2 .6 .4 .4 .2 .4 .6 .4 .2 .6 .6 = .0000005
x x x x

Figure 5 A simple probabilistic model of meter perception, using the rhythmic pattern and clocks shown in Figure 4. Assume that $P(\text{note} \mid \text{beat}) = .8$ and $P(\text{note} \mid \text{non-beat}) = .4$. Probabilities above each clock show the probabilities for each time point; $P(\text{pattern} \mid \text{clock})$ is the product of these probabilities. The model correctly predicts the preference for clock (a) over clock (b).

Another highly influential line of research in recent meter-finding has been the work of Large and colleagues, using the approach of dynamical systems (Large & Jones, 1999; Large & Kolen, 1994). By this view, meter perception involves an oscillator—a mechanism that generates a periodic wave; given an input pattern, the oscillator gradually (though perhaps quite quickly) adapts to the period and phase of the input, so that the peaks of the wave coincide with events in the input. The oscillator approach has a number of attractive properties. Once set in motion by external events, the oscillator naturally continues, its peaks representing expectations for future events. If events in the input are highly regular, the oscillator's peaks are sharp and high, representing very strong and specific expectations; if the input is less regular, a flatter response results. Adding additional oscillators can incorporate multiple metrical levels. Figure 6 shows the entrainment of an oscillator to an input pattern, as well as its generation of expectations for future events. Oscillator models are supported by experiments showing that the identity between two time intervals is more accurately judged when presented in the context of a regular pulse of similar intervals (Large & Jones, 1999) and that events occurring at expected positions (in relation to a previously established pattern) are more accurately judged with respect to pitch (Jones, Moynihan, MacKenzie, & Puente, 2002). Other researchers have also adopted the oscillator approach to meter-finding, including McAuley (1995), Gasser, Eck, and Port (1999), and Toiviainen (2001).

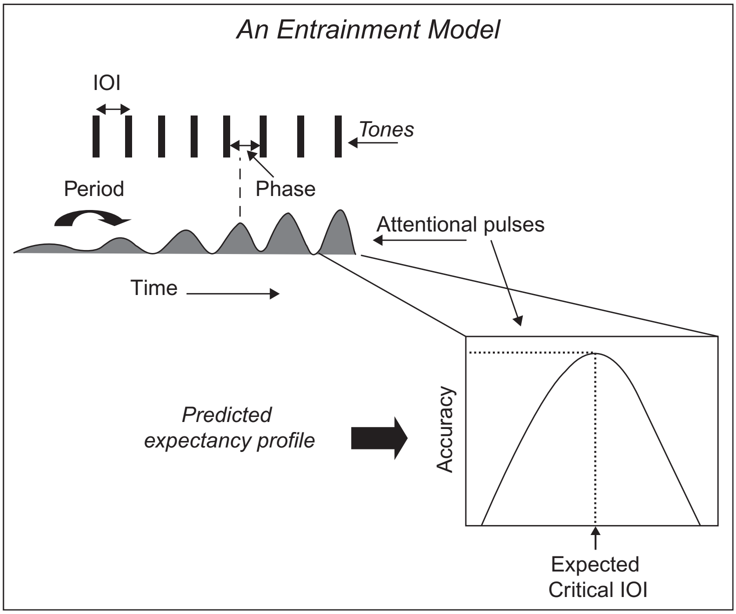

Figure 6 An oscillator entraining to a regular input pattern. Peaks in the oscillator's pulses represent expectations for future events.
From Jones et al. (2002). ©2002 American Psychological Society. Used by permission.

In recent years, much work has been devoted to identifying meter in audio input. A pioneering study in this regard was that of Scheirer (1998). Scheirer's model divides the input into frequency bands, differentiates the amplitude of each band, and finds spikes where the energy is greatly increasing; this input is then fed to "comb filters," each of which resonates to energy of a certain period. Another widely used approach in audio models is autocorrelation, which searches for self-similarities in the audio signal over different time periods and uses this to infer metrical levels (Goto, 2001; Gouyon & Herrera, 2003). Most work on meter-finding in audio appears to be more concerned with practical information-processing problems than with modeling cognition, thus we will not consider it further here; for a survey of this work, see Gouyon and Dixon (2005).

Despite the achievements of work on meter-finding during the past four decades, important challenges remain. Most meter-finding models are guided (either explicitly or implicitly) by a few simple rhythmic principles, notably the preference to align notes (especially long notes) with strong beats. But research in music theory (e.g., Lerdahl & Jackendoff, 1983) has identified a number of other factors in meter perception that, although subtle, can clearly play a role in some circumstances. Consider, once again, the melody in Figure 3b. The correct metrical analysis (shown above the staff) seems clear enough; yet few, if any, meter models could obtain this result, since the notes are all durationally equal except for the last note of each phrase. (Some models would consider the last note of each phrase to be metrically strong, because they are longest; this inference is of course incorrect in this case.) If we consider *why* we hear the meter we do, a crucial factor would appear to be parallelism—the fact that the melody contains a repeated three-note pattern (marked with brackets above the score); this strongly favors a metrical level with beats located at the same place in each instance of the pattern. However, parallelism indicates only the period of the meter (the time interval between beats), not the phase (the location of the beats). Why we tend to hear the third note of the pattern as strong, rather than the first or second, is not obvious; it may be because it is left by leap and therefore has more structural importance. In fairness, the melody is perhaps somewhat ambiguous with regard to phase: it is not difficult to hear it with the first and last notes of each phrase metrically strong, as in Figure 3c. Pitch may also have subtle effects on meter perception in other ways, for example, through the influence of harmony: there is a preference for changes of harmony on strong beats. (In the second phrase of Figure 3b, this reinforces the notated meter, as there is an implied move from I to V7 on the second to last note.) Incorporating the influence of pitch will be a major challenge for future research in meter-finding.

IV. Other Aspects of Perception

Although key-finding and meter-finding have been the most widely studied problems in computational music modeling, a number of other problems have also received attention. What follows is a survey of some of these problems and approaches that have been applied to them.

A. Pitch Identification

How does the human auditory system convert the incoming acoustic signal into a pattern of pitches? This is perhaps the most basic question in music cognition, though its importance extends well beyond music. The problem has been widely studied from a variety of perspectives but is still not fully solved; it is treated only briefly here (see Oxenham, Chapter 1, and Deutsch, Chapter 6, for detailed discussions). Many models of pitch perception have come from auditory psychology, often building on evidence about the physiology and neurology of the auditory system (see de Cheveigne, 2005, for a review). Other work comes from the engineering side, where the problem is generally known as "automatic transcription" (see Klapuri, 2004, for a review). Most of the studies in this latter category have practical information-retrieval purposes in mind and do not purport to model human pitch perception. Yet the two lines of research have followed strikingly convergent paths, encountering similar problems and often finding similar solutions to them.

At the broadest level, models of pitch perception can be categorized into those that work from a spectral analysis of the waveform (showing the energy level of the frequency components that comprise the waveform) and those that work from the waveform itself. With regard to the spectral approach, the simplest method of pitch identification is to take the frequency of highest energy, but this often does not work; a pitch can be perceived even in a signal where the corresponding frequency is entirely absent (the case of the "missing fundamental"). A more robust approach is to give "credit" to each possible fundamental pitch for each frequency that is a multiple of it (and thus a possible harmonic). This is essentially equivalent to creating a spectrum of the spectrum, known as a *cepstrum*, and identifying pitches at the peaks in the cepstrum. Another strategy is to take intervals between spectral peaks as indicators of underlying fundamental frequencies. Turning to the waveform approach, the underlying idea here is that a waveform generally reveals a strong periodicity at the underlying pitch, even when the fundamental is not present. Autocorrelation can be used to identify the time intervals at which repetition is strongest. (Mathematically, this is similar to cepstral analysis.) Klapuri (2004) and de Cheveigne (2005) explore the advantages and disadvantages of these various approaches, with regard to both their practical efficacy and their physiological and psychological plausibility.

The problem of pitch perception is further complicated when we consider polyphonic music, in which more than one fundamental is present at a time. In this case, the problem becomes one of grouping partials (frequency components) together in the correct way. Of great importance here is the fact that partials belonging to the same note generally change over time in similar ways, with regard to start and stop times, changes in amplitude, amplitude modulation (tremolo), and frequency modulation (vibrato) (Rosenthal & Okuno, 1998); heuristics based on these principles have proven to be very useful in transcription. In addition, “top-down” musical knowledge (such as expectations about what pitches are likely to occur and when) can be brought to bear on the transcription process; Bayesian probabilistic methods are of particular value in this regard (Kashino, Nakadai, Kinoshita, & Tanaka, 1998).

A related problem—one that also assumes audio rather than symbolic input—is the perception of timbre. Few studies of timbre have proposed computational models, as the term is defined here. Some proposals might be included in this category, however, such as spatial representations of timbral similarity (Grey, 1977) and systems for timbral categorization of sounds (Fujinaga, 1998). See McAdams, Chapter 2, for a detailed discussion of timbre.

B. Grouping and Voice Separation

In listening to a melody, we usually group the notes into short temporal chunks—motives and phrases—and then into larger segments such as periods and sections. A number of computational models of melodic grouping have been proposed. Tenney and Polansky (1980) propose a simple but effective model, based on the idea that phrase boundaries tend to coincide with large intervals in some musical dimension—either large temporal intervals (rests or long notes), large pitch intervals, or even intervals (changes) in some other dimension such as dynamics or timbre. In Tenney and Polansky’s model, each melodic interval is assigned a “distance,” defined as a weighted combination of the interval sizes in all four of these dimensions. Intervals whose distances are local maxima—greater than the distances on either side—are defined as grouping boundaries, and then higher-level groups are formed from smaller ones by the same principle. The model is designed for 20th-century art music and is tested on several monophonic 20th-century pieces; in this respect it stands apart from other grouping models and indeed computational music-perception models generally, most of which are designed for Western classical music (or related styles, such as European folk music).

Several other grouping models deserve mention. The Local Boundary Detection Model of Cambouropoulos (1997) adopts an approach similar to Tenney and Polansky’s, choosing certain intervals as phrase boundaries based on their magnitude relative to other nearby intervals. More recently Cambouropoulos (2006) has incorporated repetition into the model, under the reasoning that we favor phrase boundaries that align with repeated segments. The theory of Lerdahl and Jackendoff (1983), which proposes a set of criteria or “preference rules” that determine grouping boundaries, is the basis for the computational models of Temperley (2001) and Frankland and Cohen (2004). Finally, Bod’s probabilistic model (2002) represents a melody as a simple tree structure, dividing into phrases and then into notes; the model incorporates statistical knowledge about the number of phrases per melody, the number of notes per phrase, and the likely positions of different scale degrees in the phrase, and uses that knowledge to infer the most likely phrase boundaries.

Virtually all computational models of grouping have been confined to monophonic music—a major oversimplification of musical experience, since most of the music we hear is polyphonic. However, recent experimental work (Bruderer, McKinney, & Kohlrausch, 2010) suggests that people tend to segment a piece in much the same way whether they are given just the melody or a full polyphonic texture. A plausible hypothesis about polyphonic grouping, then, is that listeners first extract the melody, segment it, and then impose that grouping on the full texture. Most computational models have also been somewhat limited in that they have addressed only a single low level of grouping—roughly speaking, the level of the phrase. As noted earlier, grouping is usually assumed to be hierarchical, with smaller units combining into larger ones; large sections of an extended piece might be several minutes in length. Clearly, the modeling of high-level segmentation would be a formidable challenge, requiring sophisticated analysis of tonal and thematic structure and knowledge of formal conventions.

Listening to polyphonic music also involves grouping notes into lines or voices. This, too, presents a challenging problem for computational modeling—one known by various names such as voice separation, stream segregation, and contrapuntal analysis. Gjerdingen (1994) proposes a connectionist model of voice separation in which sounding pitches activate units in a two-dimensional pitch-time array; the activation spreads to nearby units in pitch and time, such that the maximally activated units between two pitches form a line connecting them. Marsden's model (1992) begins by creating links (potential linear connections) between nearby notes, with the weight of a link depending on the corresponding pitch interval; links compete with one another and form networks of compatible links, in a manner similar to a connectionist network. Similar in spirit is the "predicate" model of Kirlin and Utgoff (2005); this model makes a series of decisions as to whether a pair of notes belong in the same voice or not (depending on their pitch proximity, temporal distance, and other factors), and then inductively assigns each note to the same voice as the previous note that it is joined with. Temperley (2001) proposes an optimization model of voice separation in which a large set of possible analyses are considered and evaluated by several criteria; a "good" analysis is one in which there are relatively few streams and relatively few large leaps and long rests within streams. Finally, the model of Kilian and Hoos (2002) is similar to an optimization model and uses criteria similar to Temperley's to evaluate analyses; rather than searching exhaustively for the best analysis, however, it begins with an initial analysis and randomly alters it, keeping alterations that are beneficial according to the evaluation function. An unusual feature of Kilian and Hoos's model is that it allows multiple simultaneous notes within a single voice.

C. Harmonic Analysis and Pitch Spelling

Another aspect of music perception that has been explored in computational modeling is harmonic analysis—the identification of harmonies in a pattern of pitches. At a minimum, a harmonic analysis model must identify roots; other information may also be extracted, such as chord quality (major versus minor, triad versus seventh) and "Roman numeral" labels (which requires knowledge of the key as well). Harmonic analysis requires not only the assignment of labels to segments but also the demarcation of the segments themselves, that is, the identification of points of harmonic change. The problem is further complicated by nonchord tones (notes that are not part of the current chord), implied and incomplete chords (in some cases, a harmony may be implied by a single note), and the role of meter and other contextual factors.

Early attempts at computational harmonic analysis were put forth by Winograd (1968) and Maxwell (1992). These ambitious models generate complete Roman numeral analyses, indicating both harmony and key. They are, however, extremely complex, involving elaborate procedures and a large amount of built-in knowledge (such as knowledge about common patterns of harmonic progression). At the opposite extreme in this regard is Parncutt's psychoacoustic model of harmony (1989), building on Terhardt's virtual pitch theory (1974). Parncutt's model takes acoustic input and operates on a single simple principle: the root of a chord is the pitch most strongly implied by the combined partials of all the notes. The model offers an elegant way of judging the roots of isolated chords and generally yields quite successful results at this task; however, it provides no solution to the segmentation problem and the other complexities of harmonic analysis just mentioned, such as nonchord tones and implied harmonies.

Several recent studies of harmonic analysis have proposed optimization models. In the model of Temperley and Sleator (1999), analyses are evaluated by three main criteria: the compatibility of each pitch with the local root (generally, each note must be a chord tone of the root unless it is closely followed by stepwise motion); the alignment of changes of harmony with strong beats of the meter; and the circle-of-fifths distance between adjacent harmonies, with a preference for motions between chords that are close together on the circle. The technique of dynamic programming is used to find the best harmonic analysis of a piece out of all the possible ones. Pardo and Birmingham (2002) offer a similar approach; this model makes the simplifying assumption that changes of harmony can only occur at changes of pitch, thus greatly reducing the number of possible segment boundaries that must be considered. Raphael and Stoddard (2004) propose a probabilistic model that identifies both harmony and key, using Bayesian logic similar to that described in Section III; the probability of a harmonic structure given a note pattern depends on the probability of the harmonic structure itself (so that, for example, transitions that remain in the same key are more probable than those that do not) and on the probability of each pitch given the current key and root (this is defined by profiles similar to the key profiles described in Section II). Although the model has not been extensively tested, it holds great promise; among other things, it offers an elegant way of capturing the effects of harmony on key identification, discussed in Section II.

Connectionist methods have also been applied to harmonic analysis. Especially worthy of note in this regard is the MUSACT model of Bharucha (1987). Bharucha posits a three-level network of units representing pitches, chords, and keys; sounding pitches activate pitch nodes, pitch nodes activate nodes of chords that contain

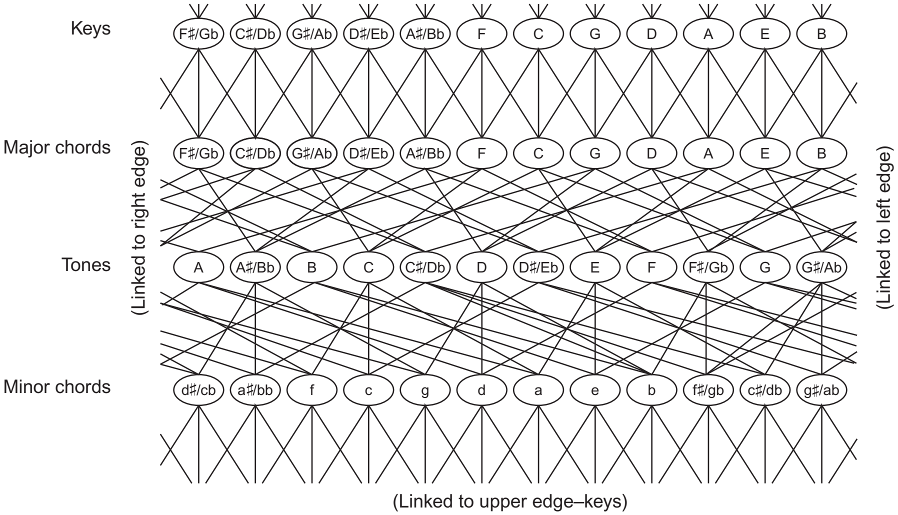

Figure 7 A network representing relationships between tones, chords, and keys. From Bharucha (1987). ©1987 The Regents of the University of California. Used by permission.

them, and chord nodes activate nodes of keys to which they belong (see Figure 7). In turn, top-down activation passes from key nodes to chord nodes and from chord nodes to pitch nodes, so that nodes for pitches and chords that have not been heard may be somewhat activated. This aspect of the model has been invoked to explain the phenomenon of "harmonic priming"—the fact that people show enhanced sensitivity to chords that are closely related to the prior context. Although the original version of the MUSACT model involved no learning, further work (Tillmann, Bharucha, & Bigand, 2000) has shown that a very similar model can be produced using the paradigm of "self-organizing maps," in which the model—initialized with random activation weights—learns to associate nodes with chords and keys in the appropriate way.

An aspect of musical structure related to harmony, though somewhat esoteric, is pitch spelling—the assignment of labels such as A♭ or G♯ to pitch events. Although there has been no experimental work on the psychological reality of pitch spelling, it is clear that trained musicians, at least, are able to choose pitch spelling labels (in transcribing a melody, for example) in an appropriate and consistent way. Several computational models of this process have been put forth, including a proposal by Longuet-Higgins and Steedman in their 1971 article (discussed earlier) as well as several more recent models (Cambouropoulos, 2003; Chew & Chen, 2005; Meredith, 2006; Temperley, 2001). Each of these models employs some kind of spatial representation, favoring spellings that locate pitches compactly in the space. A simple but highly effective principle that plays a role in several pitch-spelling models is to prefer spellings that are close together on the "line of fifths"—the circle of fifths stretched out into an infinite line (...F♯ B E A D G C F B♭ E♭ A♭ D♭ G♭...). This generally yields a satisfactory spelling of notes within the key, since the pitches of a diatonic scale form a compact set of points on the line. This criterion alone is not sufficient, however. Voice leading must also be considered: In a C-major context, for example, the chromatic pitch G♯/A♭ will be spelled as A♭ if it resolves to G, but as G♯ if it resolves to A.

Pitch spelling interacts in a complex way with harmony and also with key. Returning to the melody in Figure 1b, if the notes are rearranged as in Figure 1d, the D♯ would now probably be spelled as E♭; the harmonic implications are now altered—the opening of the melody is heard as outlining a C-minor triad rather than C major—and perhaps the key implications as well (first C minor, then shifting to E minor). The causal relationships between these processes are not obvious, however: are pitch spellings identified first, then serving as input to harmony and key, or are they only inferred once key and harmony have been determined? Computational models of pitch spelling have taken various positions on this issue. In Longuet-Higgins and Steedman's model, key is identified first and then serves as input to the pitch-spelling process; in Cambouropoulos's model, pitch spellings are identified without key information. In Meredith's model, the spelling of a pitch is theoretically dependent on the key, but key is identified in an indeterminate way; each pitch acts as a possible tonic, weighted by its frequency of occurrence in the context.

D. Pattern Discovery

A rather different kind of information extraction from those considered so far is pattern discovery—the identification of repeated themes or motives in a piece. Without doubt, pattern identification is one of the more subjective aspects of music perception. While much work in music theory and analysis has been concerned, in some way, with pattern discovery (including Schenkerian analysis, semiotic analysis, pitch-class set theory, and more traditional kinds of motivic analysis), this work has mostly been aimed more at enriching the listener's experience rather than describing it. Thus it is often unclear, in musical pattern analysis, what the "right answer" would be. Although classic experimental studies by Dowling (1978) and Deutsch (1980) identified factors involved in pattern identification, more recent research has not yielded very strong or specific conclusions (Deliège, 2001; Lamont & Dibben, 2001). Despite these uncertainties, a number of models of pattern discovery have been proposed in recent years. Most models restrict themselves to the detection of patterns within a single melodic line (Cambouropoulos, 2006; Conklin & Anagnastopoulou, 2006; Lartillot & Saint-James, 2004; Rolland, 1999), though Meredith, Lemström, and Wiggins (2002) propose a system for pattern detection in polyphonic music.

One of the challenges of pattern discovery is that musical repetition is often not exact; one instance of a pattern may have slightly different pitches or rhythms, or more or fewer notes, than another. Thus, some kind of approximate matching must be allowed. A further problem is that—even if only exact repetitions or transpositions are considered—the vast majority of repeated patterns in a piece are not ones that would be considered perceptually significant; thus some method must be found for selecting the significant ones. Meredith et al. adopt heuristics such as coverage (the number of events in the piece covered by instances the pattern) and compactness (the number of events within the temporal and registral "span" of a single instance of the pattern that are covered by it) to choose significant patterns. An alternative solution is the multiple-viewpoint approach (Conklin & Anagnastopoulou, 2006; Conklin & Bergeron, 2008), which allows a variety of dimensions of events to be considered besides pitch and duration, such as scale degree, metrical placement, and contour, thus allowing a more intelligent pattern selection process than one based on pitch and duration alone. Also of interest is Cambouropoulos's segmentation model (2006), described earlier, which identifies repeated patterns as part of the segmentation process.

E. Pitch Reduction

A final aspect of perception deserving consideration is pitch reduction: the identification of hierarchical relationships among pitch events. The status of pitch reduction with regard to music cognition is by no means obvious. The most widely practiced theory of reduction, that of Schenker (1935/1979), was not intended—and is not generally construed today—as a theory of listening, but rather, as an ideal kind of musical understanding to which listeners should aspire. Having said that, Schenkerian analysis is still a cognitive activity—at least as practiced by music theorists—and it is perfectly valid to try to model it. Reduction is extremely challenging from a modeling perspective, as it requires extensive musical knowledge (e.g., knowledge of counterpoint) and a variety of kinds of structural information about the piece, such as harmonic structure, motivic connections, phrase structure, and meter. Current implementations of Schenkerian analysis show significant progress on the problem, but are also severely limited; they either require input that is very simple (Marsden, 2010) or already somewhat reduced (Kassler, 1977; Kirlin & Utgoff, 2008), or they require guidance from the user (Smoliar, 1980).

The most sophisticated model of Schenkerian analysis to date, that of Marsden (2010), begins with a piece divided into minimal segments, each one represented by a "chord" (a set of pitches); segments are grouped together into larger ones in a recursive fashion, and at each step some notes are reduced out, leaving only the "structural" ones. The system uses chart parsing, a technique from computational linguistics that allows a huge number of possible analyses to be searched in an efficient manner. A set of heuristics helps the system choose the best analysis: for example, a preference for higher-level segments to start on strong beats, and a preference for joining together segments related by small melodic intervals. Weights for the heuristics are obtained by examining their ability to distinguish "correct" analyses (taken from the literature) from incorrect ones. The system is tested on several short Mozart themes and performs quite well; because of its computational complexity, it is currently unable to handle longer musical passages.

Also worthy of mention is Lerdahl and Jackendoff's theory of pitch reduction, presented in their book *A Generative Theory of Tonal Music* (*GTTM*) (1983). Like Schenkerian theory, the *GTTM* reduction theory requires consideration of a large number of factors, presenting a daunting computational challenge. Hamanaka and colleagues (Hamanaka, Hirata, & Tojo, 2006; Hamanaka & Tojo, 2009) present an impressive effort to model *GTTM*'s reductional component. At present, the model still requires extensive guidance from the user to yield good results, but efforts are underway to make the system fully automated.

V. Models of Musical Experience

Although extracting structural information is an important part of listening to music, there is of course more to musical experience than that; the listener must also comprehend the implications and meanings of that information. In this regard, one might think first of emotion. Much experimental research has focused on music and emotion (see Juslin and Sloboda, Chapter 15), but there is little in this work that could be considered a computational model. I will focus here on three other aspects of musical experience that have been quite extensively studied in computational research: expectation, tension, and similarity.

It has long been thought that expectation plays an important role in musical experience. A number of models have sought to predict the expectations that listeners form for subsequent events. (The vast majority of work in expectation has focused on pitch; one exception is the work of Large, Jones, and colleagues, discussed in Section III.) The most influential proposal in this regard has been the Implication-Realization (I-R) theory of Narmour (1990). Narmour's theory predicts melodic expectations on the basis of principles of melodic shape: For example, a large interval creates expectation for a change of direction, and a small interval creates expectation for another small interval in the same direction. The I-R theory in its original form is highly complex; the basic principles or "primitives" of the theory interact and combine in a variety of ways and operate at multiple hierarchical levels. Several other authors have proposed ways of simplifying and quantifying the theory (Cuddy & Lunney, 1995; Krumhansl, 1995; Schellenberg, 1997). The details of these models vary with regard to the number of factors included and the way they are defined; Figure 8 shows one proposal. Experiments have tested these models, yielding high correlations with human expectation judgments.

Narmour's theory purports to address aspects of music perception that are innate and universal; for this reason, it largely excludes considerations of harmony and tonality (though some implementations of the theory have included such factors). Other models of expectation have sought to incorporate harmonic and tonal factors. The theory of Larson (1997–1998, 2004) models expectation by analogy with physical forces: gravity favors descending over ascending intervals, inertia favors a continuation in the direction the melody has been going, and magnetism

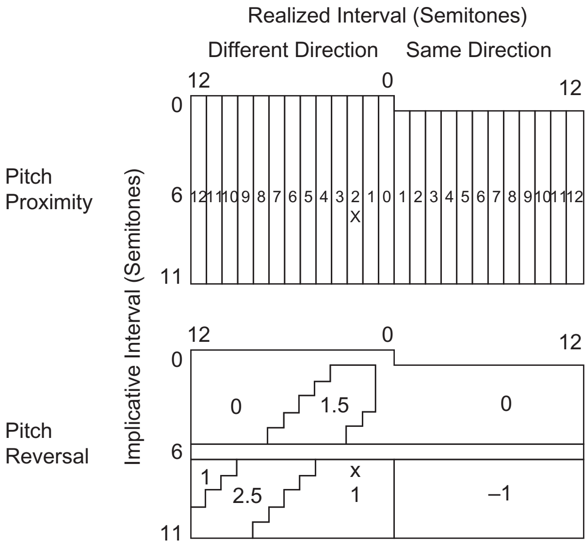

Figure 8 A two-factor quantification of Narmour's (1990) Implication-Realization model. Numbers indicate the expectedness of one ("realized") interval after another ("implicative") interval. This depends on the size of the implicative interval (indicated on the vertical axis) and the size and relative direction of the realized interval (on the horizontal axis). The expectedness of the realized interval is a weighted sum of the values for the pitch-proximity factor and the pitch-reversal factor. For example, for a realized interval of −2 semitones after an implicative interval of +7 semitones, pitch proximity yields a value of 2 and pitch reversal yields 1 (see x's).
From Schellenberg (1997). ©1997 The Regents of the University of California. Used by permission.

favors moves to pitches that are tonally more stable—for example, to a diatonic note from a chromatic one, or to a tonic-triad note from another diatonic note. An interesting feature of Larson's model is that it can produce expectations for multinote continuations; from scale degree $\hat{4}$, for example, the model predicts a move to $\hat{3}$, but it may then predict a further move to $\hat{1}$, which is more stable at a higher level.

A number of other expectation models have been proposed. Similar in spirit to Larson's model, Lerdahl's theory of tonal attraction (2001) also employs a physical metaphor, invoking gravitation but in a different way: Given a melodic context, the attraction to (expectation for) a possible following pitch is related to its harmonic stability and inversely related to the square of its distance from the current pitch. The model of Margulis (2005) combines Narmour's principles of melodic shape with the tonal factors proposed by Larson and Lerdahl. Finally, a very different approach is seen in the "multiple-viewpoint" model of Pearce and Wiggins (2006); this model considers a variety of musical parameters such as pitch interval, time interval, scale degree, and contour and generates expectations based on statistical regularities in these parameters. The multiple-viewpoint approach will be considered further later.

Another related aspect of musical experience is tension. Fluctuations in tension are an important part of music's expressive effect: It is generally agreed, for example, that a suspension is tense in relation to its resolution and that a move away from the tonic key causes tension at a higher level. Of particular importance here is the work of Lerdahl (2001). Lerdahl proposes a theory of tension building on Lerdahl and Jackendoff's theory of reduction (1983) and his more recent theory of pitch space. (Lerdahl's pitch-space theory posits a spatial representation of relations between keys and chords, thus capturing intuitions about the "distance" of harmonic motions—an important aspect of musical experience in and of itself.) In Lerdahl's theory, the "hierarchical tension" of an event is related to its degree of embedding in a reductional tree; more deeply embedded events are higher in tension (see Figure 9). Hierarchical tension is also influenced by the distance in pitch space between each event and the event that it elaborates (also considering the distance between the parent event and *its* parent, and so on recursively up the tree). Thus, a chromatic chord in the middle of a phrase is likely to be high in tension both because it is deeply embedded in the tree and because it is harmonically distant from superordinate events. The model also considers sequential tension (related to the pitch-space distance between each event and the next) and dissonance. Recent experiments have found strong correlations between the model's predictions and human tension judgments (Lerdahl & Krumhansl, 2007), although it appears also that perceived tension depends primarily on local context (e.g., the immediate phrase) rather than on large-scale tonal structure (Bigand & Parncutt, 1999).

We conclude this section with an important but elusive aspect of musical experience: similarity. A number of studies have sought to model intuitions about musical similarity, mostly focusing on melodies. Musical similarity is a very broad concept, however, and can mean a variety of things. One kind of similarity, which we might call "global similarity," is concerned with general properties of pieces such as style, mood, or rhythmic feel. Studies of this kind of similarity often employ a multiple regression approach, examining a number of dimensions or features of melodies and the value of each one in predicting similarity judgments (Eerola, Järvinen, Louhivuori, & Toiviainen, 2001; Monahan & Carterette, 1985).

Another kind of musical similarity, which we might call "pattern similarity," is concerned with specific patterns of pitch and rhythm; this is the kind of similarity that can make one melody seem like a variant of another. Various statistical methods have been proposed for evaluating pattern similarity. Schmuckler (1999) applies Fourier analysis to pitch patterns, revealing underlying pitch contours at different time scales. Juhasz (2000) divides pitch patterns into small time slices and represents them in a high-dimensional space; the dimensions of greatest variance in the space then reveal "general melodic designs" that that can be used to categorize melodies. Detecting pattern similarity is similar to the problem of

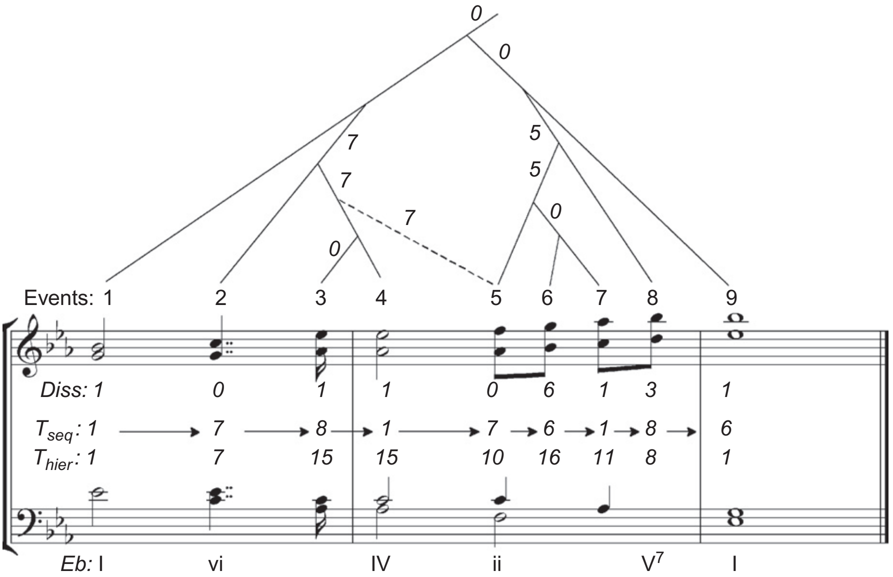
0
0
7
7
7
0
5
5
0
Events: 1 2 3 4 5 6 7 8 9
Diss: 1 0 1 1 0 6 1 3 1
Tseq: 1 → 7 → 8 → 1 → 7 → 6 → 1 → 8 → 6
Thier: 1 7 15 15 10 16 11 8 1
Eb: I vi IV ii V7 I

Figure 9 The predictions of Lerdahl's (2001) tension model for a phrase from Wagner's *Parsifal*. Tension is predicted as a function of hierarchical tension (T_{hier}) and sequential tension (T_{seq}). The tree above the staff indicates prolongational structure; numbers on the branches indicate pitch-space distances between events. The hierarchical tension of an event is the sum of the distance values from the event to the top of the tree, plus its surface dissonance value (*Diss.*): for example, for the first chord of the second measure, $T_{hier} = 7 + 7 + 1 = 15$.

From Lerdahl and Krumhansl (2007). ©2007 The Regents of the University of California. Used by permission.

detecting repeated patterns within a piece, discussed earlier in Section IV; Rolland (1999) bridges these two problems, using a pattern discovery system to identify repeated pitch sequences both within and between jazz solos.

Some studies of pattern similarity focus on practical applications more than on cognitive modeling (Chai, 2004; Pardo, Shifrin, & Birmingham, 2004; Typke, Giannopoulos, Veltkamp, Wiering, & van Oostrum, 2003); a particular concern has been the development of searchable musical databases, a problem sometimes known as "query-by-humming." Nonetheless, these studies offer a number of ideas that may be relevant to the modeling of human similarity judgments as well. As Pardo et al. point out, assessing the similarity between one melody and another involves finding the best way of mapping one onto the other. They accomplish this using a "string-matching" technique, which finds the best alignment between two note sequences, factoring in penalties for deletions and additions. Typke et al. adopt a somewhat similar approach, but assign a weight to each note in a melody, reflecting the importance of finding a match for it; this recognizes that some notes (e.g., notes in metrically strong positions) may be more essential to the identity of a melody than others. Müllensiefen and Frieler (2004) offer an interesting examination of pattern similarity measures from a psychological point of view.

VI. Models of Performance

To date, the preponderance of work in musical cognitive modeling has focused on issues of perception; in this way, it reflects the field of music cognition as a whole. There has, however, been significant work on the modeling of musical behavior as well, notably, music performance. One advantage of studying performance is that, because it is a kind of behavior, it naturally yields concrete data that can be studied and modeled (unlike perceptual processes, which do not usually generate data unless it is especially elicited, e.g., in a listening experiment). Most research on performance has focused on the piano (or other keyboard instruments). There are good practical reasons for this; because a piano performance can be completely encoded with just a few variables—the onset time, duration, and velocity (i.e., loudness) of each note—it presents a relatively manageable object of study.

One kind of performance data that lends itself well to computational modeling is performance errors. Performance errors are of interest not just in their own right, but because they provide a window into how performance actions are planned and executed. Particularly noteworthy in this regard is the range model, proposed by Palmer and Pfordresher (2003; see also Pfordresher, Palmer, & Jungers, 2007). Palmer and Pfordresher focus on movement errors, in which a note indicated in the score is played somewhere other than its correct location. The likelihood of a note being played at a given location is posited to be related to its cognitive activation, which in turn is the product of two variables, M and S. S is related to the distance between the "source" location (the correct location of the note) and the current location, and it decreases as the source location gets further away; M is the metrical similarity of the source location to the current location—the intuition being that a performer is more likely to confuse the current location with one that is metrically similar (e.g., replacing a strong-beat note with another strong-beat note). Experiments with expert and novice pianists provided support for both the distance and similarity components of the model (see Figure 10). The model also posits an effect of tempo on errors: as tempo increases, it is predicted that errors will become more frequent (the well-known "speed-accuracy trade-off"), and also that the distance between source locations and the current location (measured in "score time" rather than absolute time) will decrease; this is because a faster tempo increases cognitive demands and therefore decreases the size of the context that can be considered in planning.

Another issue that has received some attention in computational modeling research is fingering. On some instruments, notably the piano and guitar, the performer has multiple options as to which finger to use in producing a given note

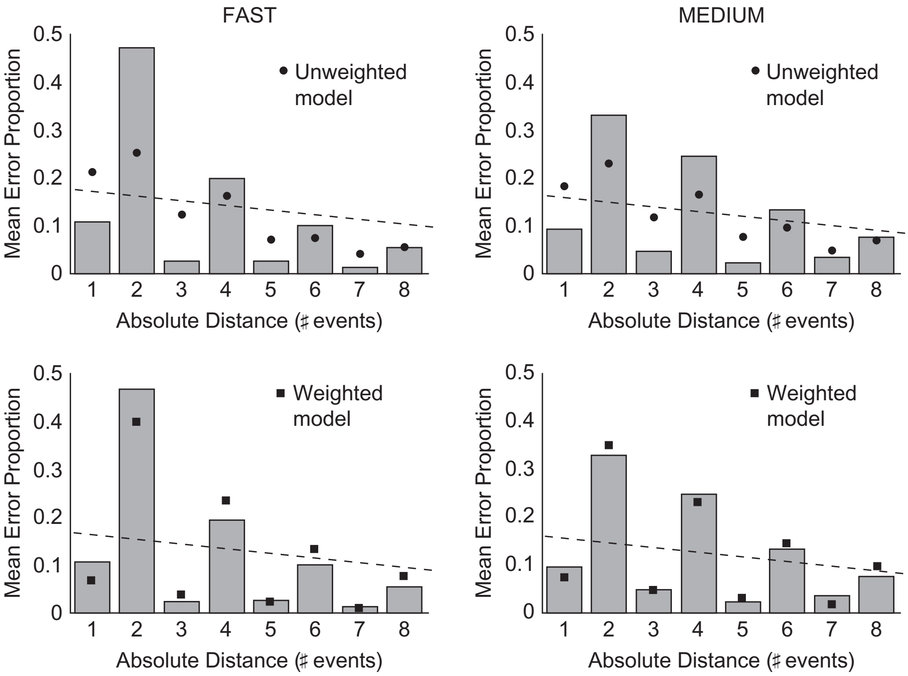

Figure 10 Mean error proportions and model fits by sequence distance and tempo conditions (left graphs: fast; right graphs: medium) for Experiment 1. The fact that error rates are higher for even-numbered distances shows that performers are more likely to confuse a note with another note in a similar metrical position. Data is shown for the unweighted model, which assigns a constant degree of difference to each pair of adjacent metrical levels, and the weighted model, which assigns varying degrees of difference. The dashed lines indicate chance estimates.
From Palmer and Pfordresher (2003). ©2003 American Psychological Association.

(on guitar, there may also be a choice of strings); finding efficient fingering patterns is an important part of skilled performance on these instruments. Parncutt, Sloboda, Clarke, Raekallio, and Desain (1997) propose an optimization model of piano fingering. Given a short sequence of notes, the model assigns a score for each possible fingering, factoring in a set of preferences—for example, prefer for note-to-note intervals to match the natural distance between the fingers used for them, prefer to avoid changes of hand position, and prefer to avoid placing the thumb on a black note. Violations of these preferences incur penalties, and the best fingering is the one incurring the minimum score. Dynamic programming is used to find the optimal fingering pattern for longer passages. Studies of guitar fingering by Sayegh (1989) and Radicioni and Lombardo (2005) adopt a similar approach: local constraints and preferences are defined for individual chord fingerings, and the goal is to find the optimal sequence of fingerings for a series of chords. As

Sayegh points out, an added complication in the case of guitar fingering is the fact that the same note played on different strings has a different *sound*, and this too must be taken into account.

Notwithstanding occasional attention to problems such as performance errors and fingering, by far the largest share of computational research on music performance has focused on the issue of performance expression. Performance expression refers to the intentional manipulation of musical performance parameters for expressive purposes, beyond what is specified in the score: in particular, nuances of timing, dynamics, and articulation (and, on some instruments, other parameters such as pitch, vibrato, and timbre). Given the breadth and importance of computational modeling work in this area, it will be surveyed in somewhat greater depth.

One of the earliest and most well-developed models of performance expression is the Director Musices (or KTH) model (Friberg, Bresin, & Sundberg, 2006; Sundberg, 1988). This model relates performance dimensions—primarily timing and dynamics, but other dimensions as well—to aspects of musical structure. An initial set of rules, derived through trial-and-error experimentation, included (1) increasing volume with higher pitch, (2) shortening very short notes (in relation to their notated length), and (3) slowing down and increasing dynamics at points of high "harmonic charge," where the harmonic charge of an event depends on its harmonic distance from the tonic on the circle of fifths. (The latter rule is of particular interest, as it is one of the few attempts to relate performance expression to harmony.) More recent work has developed this approach through experimental testing, the addition of rules pertaining to other musical parameters such as intonation and articulation, and exploration of further issues such as the communication of emotion.

The Director Musices model also postulates a decrease in tempo and dynamics at the ends of phrases—a very well-established principle in music performance research. Todd's model of expressive performance (N. Todd, 1989, 1992) also incorporates this principle; the innovation of the model is to apply it in a hierarchical fashion. (Several versions of the model have been proposed; only the latest [1992] version will be discussed here.) A hierarchical grouping structure of phrases and sections is used as input; each unit of the phrase structure generates a "V"-shaped profile. These profiles sum together over levels to produce a single curve that controls both timing and dynamics. This produces the intuitively satisfying result that tempo deceleration is greater at the ends of large segments (i.e., sections) than smaller ones (phrases). Figure 11 shows data from an experimental test of the model; it can be seen that the model does indeed capture large-scale features of the performance, though many local fluctations in timing are evident as well. Another approach to expressive timing is seen in Mazzola's *Rubato* model (Mazzola & Zahorka, 1994; Müller & Mazzola, 2003). Mazzola views tempo itself in a hierarchical way, such that each piece has a main tempo, sections within the piece may have subordinate tempi, and moments within each section (such as an appoggiatura or the end of a phrase) may have still lower-level tempi; the result is a complex timing profile somewhat similar to those produced by Todd's model.

Given the interest throughout cognitive science in machine learning—systems whose parameters and rules can be learned from data—it is not surprising that this

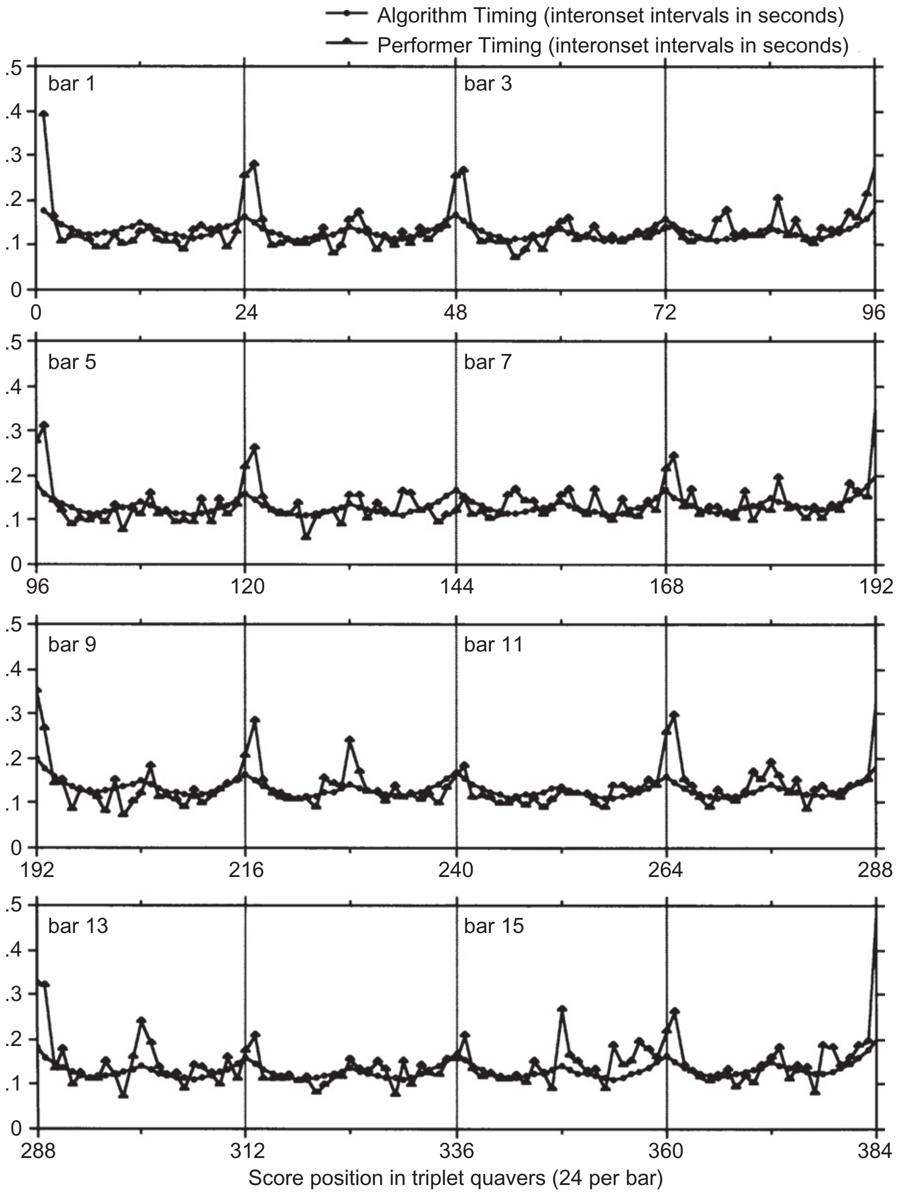

Figure 11 An experimental test of Todd's (1992) model of expressive timing (using a Schubert Impromptu). The smooth line indicates predictions of the model; the jagged line indicates the timing of a human performer.
From Windsor and Clarke (1997). ©1997 The Regents of the University of California. Used by permission.

approach has been applied in modeling music performance as well. A case in point is the work of Widmer and colleagues (Widmer, 2002; Widmer & Tobudic, 2003). In their approach, statistical analyses of human performances search for correlations between performance features (such as increases or decreases in tempo or dynamics) and features of musical structure: note length, melodic shape, harmonic progression, and the like. This process yields a number of quite specific rules, such as “given two notes of equal length followed by a third longer note, lengthen the second note.” In a variant of this approach, a large number of performances are stored in memory; when a new passage is encountered, the model searches for the passage in the memory bank that is most similar to it in terms of musical features, and the timing and dynamics of that passage are applied to the new one.

Several other models of expressive performance deserve brief mention. Bresin (1998) applies the approach of neural-network modeling; trained on data from human performances, the network learns to generate timing patterns from input parameters such as pitch interval and phrase position. Dannenberg and Derenyi (1998) incorporate expressive parameters into a trumpet synthesis model—one of the few attempts to address performance expression on instruments other than piano. Raphael’s Music Plus One system (2001) addresses the related problem of expressive accompaniment: the system can play along with a human performer in real time. This requires not only generating a performance that is musical and expressive in its own right, but also synchronizing that performance with the other player; the system learns from previous performances by the player, allowing it to predict the timing and dynamics of each note.

VII. Models of Composition

The application of computers to problems of composition has a long history; indeed, some of the earliest musical uses of computers were of this nature (Hiller & Isaacson, 1959; Pinkerton, 1956). Here again, however, it is important to consider the issue of purpose. In many cases, the use of computers in composition is for practical purposes: they may be used to generate timbres that could not be generated in other ways, or programmed to make stochastic (random) choices with the aim of producing results *different* from those of a human composer. In other cases, however, computers are used with the aim of simulating human composition, thus shedding light on psychological processes; such applications deserve consideration here.

Much computer-aided composition has been guided by a single unifying principle: a musical style consists of various kinds of events in some statistical distribution, and the style may be simulated—at least in an approximate way—by producing new music with the same distribution. Normally, the “events” in question are not simply pitches or durations considered in isolation but also take context into account in some way. The simplest and most widely used example is the Markov chain, a model that represents the probability of each element (scale degrees, for example) given a context of a specified size. A zeroth-order Markov chain simply represents the probability of each element in isolation; a first-order model indicates the probability of each element following another single element (for example, the probability of scale degree $\hat{2}$ occurring after scale degree $\hat{1}$); a second-order model assumes a context of two elements, and so on. This data can be extracted from a musical corpus and then used in a stochastic fashion to generate new music that reflects the same statistical properties as the corpus. Some of the earliest computational composition systems were simple applications of Markov chains, such as systems for composing nursery tunes (Pinkerton, 1956) and hymn tunes (Brooks, Hopkins, Neumann, & Wright, 1957); more recently, Ponsford, Wiggins, and Mellish (1999) have used third- and fourth-order Markov chains to generate Baroque sarabandes.

An extension of the Markov chain idea is the multiple-viewpoint approach, developed by Conklin and colleagues (Conklin & Witten, 1995; Pearce & Wiggins, 2004). Under this approach, events are represented in a variety of dimensions—for example, pitch, melodic interval, scale degree, duration, and metrical position; each of these dimensions, and each combination of them, can be the basis for a Markov chain (of zeroth or higher order). To find the best possible way of doing this, the concept of cross-entropy is used. Each possible "viewpoint" (i.e., each combination of basic dimensions) can be used to assign some probability to the training data, and the best model is the one that yields the highest probability. This model can then be used to generate new music. (The multiple-viewpoint approach has also been applied to the modeling of improvisation; see Pachet, 2002.)

Another variant on the Markov chain is the hidden Markov model, which posits an underlying network of states; output symbols are generated by transitions from one state to another. An especially impressive application of this technique is Mavromatis's system for the improvisation of Greek Orthodox Church chant (2005, 2009). Figure 12 shows a preliminary version of the model, which simply generates a sequence of pitches. Greek chant improvisation is also constrained by the text being sung, and especially the stress pattern of the text; Mavromatis's complete model takes this into account, taking a stress pattern as input and generating an appropriate melody for it. Using a technique known as Minimum Description Length, Mavromatis's system incrementally adjusts itself to find the best model, considering both simplicity (the number of nodes in the network) and goodness of fit to the training data.

A rather different technique for modeling composition, though related, is the use of neural networks. Research in the 1980s showed that neural networks could be used in a "sequential" or "recurrent" fashion to simulate processes over time—for example, predicting the next event in a sequence. Several researchers have applied this technique to composition (Franklin, 2006; Mozer, 1994; P. Todd, 1989). A network is trained on a corpus of music; then, given a starting point, its predictions can be used to construct a new piece in the style. There is a strong connection here with the Markov-chain approach: the network is essentially selecting events that have a high probability of occurring in that context. One difference, as Mozer observes (1994, pp. 2–3), is that a neural network is—in theory anyway—capable of shifting between different sizes of context (first-order, second-order, etc.) in a

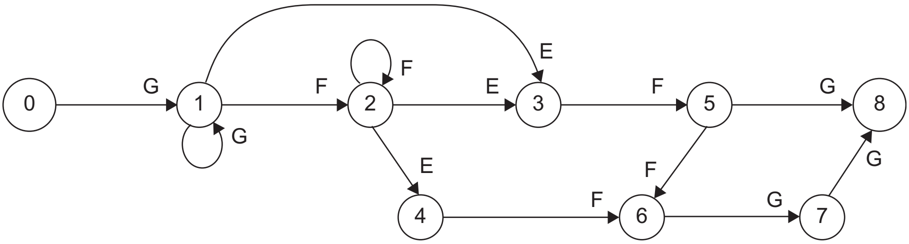

Figure 12 A hidden Markov model for generating cadential formulae for Greek church chants. Circles represent nodes (states); arrows represent possible transitions between them. (Numbers on the nodes are arbitrary and do not necessarily indicate sequential order.) Letters represent the notes that would be generated by transitions from one state to another. Transitions are assigned probabilities (not shown here), so that the probabilities of all transitions coming out of a state sum to 1.
From Mavromatis (2005). Used by permission.

dynamic fashion and finding the one that captures the most structure in the data. Todd's model exhibits interesting "creative" behaviors, such as splicing together segments of a melody in novel ways and creating a new hybrid melody out of two melodies seen in training.

Other computational composition systems employ a more traditional "rule-based" approach, sometimes in combination with stochastic processes. Rule-based strategies are seen in Baroni and Jacobini's system for generating chorale melodies (1978), Rothgeb's system for harmonizing unfigured basses (1980), and Ebcioglu's system for harmonizing chorale melodies (1988). Hiller and Isaacson's pioneering experiments (1959) included an attempt to simulate the style of Palestrina; Markov processes were used to generate material, but this was then filtered through traditional contrapuntal rules to find good solutions—an approach known as "generate-and-test." The generate-and-test approach is reflected also in the approach of genetic algorithms. Under this approach, randomly generated melodic segments are evaluated with the aid of a "fitness function"; the better segments are split apart and recombined to make new segments (sometimes with small random changes or "mutations"). The fitness function can be a human supervisor (Biles, 1994) or an automatic function (Özcan & Erçal, 2008).

Perhaps the most well-known figure in computer-aided composition is David Cope (2000, 2005), who has explored a wide range of techniques for simulating compositional styles, such as Bach chorales and Beethoven symphonic movements. Many of Cope's projects involve what he calls "recombinancy" (or "recombinance"): the idea that effective composition within a style largely consists of finding common, characteristic patterns in existing music and recombining them in new ways. Early versions of his system concatenated patterns together in a simple random fashion; later versions do so more intelligently, ensuring that patterns are used in contexts similar to where they occurred originally. Cope has also experimented with various kinds of rule-based filters to refine his system's output and with the use of hierarchical analysis to find (and then reuse) structural patterns beneath the musical surface. Although Cope's systems often produce quite impressive pieces, his explanations of them tend to be somewhat vague; crucial details are omitted, such as the exact music on which the system was trained, the size of the units that it "recombined," and the role of the human user in guiding the process and selecting the output. This makes it difficult to fully understand his models, or to evaluate them.

VIII. Conclusions

As explained at the beginning of this chapter, computational modeling research is concerned with the development of precise, testable descriptions of mental processes and representations. As such, computational modeling is a crucial part of the field of music cognition: it is this research that provides concrete accounts of—or at least, hypotheses about—the cognitive mechanisms underlying psychological data. For the most part, current cognitive models fall short of total success. But even their failures can be instructive; for example, if a meter-finding or key-finding model uses certain criteria but fails to produce adequate results, this may draw our attention to other criteria that are used in human processing and that need to be incorporated into future models.

Surveying the present state of computational music cognition, several major challenges for the future are apparent. One challenge that arises frequently is the wide range of kinds of knowledge and information that must be brought to bear. In the case of meter-finding, for example, a successful model must not only consider the time points of events (as all models do) but also pitch, harmony, motivic structure (i.e., parallelism), phrase structure, texture, and other factors as well. (Complicating the matter further, some of these structures—such as harmony and motivic structure—also require meter as input, creating a difficult "chicken-and-egg" problem.) The same point arises in a different way with models of composition. Models that simply consider low-level statistical features (e.g., transitional probabilities between notes) may be able to create music that sounds passable on a casual hearing, and such low-level features may indeed play *some* role in the human compositional process. But surely they are not the whole story; there seems little doubt that composition is also informed by knowledge about higher-level structures such as harmony, phrase structure, motivic/thematic structure, and form, and any truly plausible cognitive model of the compositional process will need to find ways of incorporating this knowledge.

Another challenge for computational music cognition—related to the first—is to find models and principles that are applicable to a range of cognitive processes. Much of the work considered in this chapter focuses quite narrowly on a single process, such as key-finding or performance expression. There is nothing wrong with this approach. But if one can find cognitive models that explain a wider range of phenomena, this has the obvious benefit of parsimony—one model is simpler than several. To take an exemplary instance of such a "multipurpose" model, Krumhansl's key-profile model (1990), originally designed to explain intuitions about the stability or "fit" of pitches in a tonal context, has been used to explain key-finding, key relations, chord relations, melodic expectation, and distributions of pitch classes in compositions—truly an impressive achievement for such a simple model. An example of a very different kind is the use of Bayesian probabilistic reasoning. As described earlier, the Bayesian approach can be used to model various kinds of "structure-finding" processes—identifying the best (most probable) structure given the surface (see Eq. 1). This approach has been applied to a number of problems in music cognition, including key-finding, meter-finding, harmonic analysis, and phrase structure analysis (see Sections II, III, and IV). But Bayesian reasoning can also be used in a very different way—to evaluate models themselves with regard to their fit to a body of data. Replacing "structure" in Equation 1 with "model," and "surface" with "data":

$$P(\text{model} \mid \text{data}) \propto P(\text{data} \mid \text{model}) \times P(\text{model}) \qquad \text{(Eq. 3)}$$

This principle can be used in models of composition, the reasoning being that the model that assigns highest probability to the data is the most plausible model of the process that gave rise to it; this is reflected in the multiple viewpoint approach (Conklin & Witten, 1995) and in the Greek chant model of Mavromatis (2005, 2009). The same principle might also be used to characterize the listener's mental representation of a musical style and thus to model processes such as melodic expectation; this is seen, for example, in the model of Pearce and Wiggins (2006). Here, then, is a simple principle that appears to have explanatory relevance to a variety of phenomena in music cognition.

A third challenge for computational models of music concerns evaluation. The most obvious way of evaluating a computational model is by examining its performance on the job that it is supposed to do. For example, one can test a key-finding model by observing the proportion of pieces (or sections of pieces) on which it chooses the correct key; one can test a model of expressive timing by comparing its "performances" to those of humans. Most recent computational modeling studies present some kind of evaluation. But the choice of test materials is of great importance: ideally, a test set should be fairly large, systematically selected (as opposed to simply taking a few pieces one happens to have handy), and broadly representative of the style of music under consideration. In addition, it is especially desirable to use test materials that have been used in previous studies, to allow comparison with other models. (In cases where a model is trained on data, it is also important that a model not be tested—at least, not exclusively—on the same data on which it was trained.) At present, very few evaluations in computational modeling studies meet all of these criteria. In this regard, the "engineering" side of computational music research offers an admirable example: The annual MIREX competitions (held in conjunction with meetings of the International Society for Music Information Retrieval) allow for the comparative testing of models on systematically selected test materials (Downie, 2006).

Finally, we must return, one last time, to the issue of purpose. Many computational music studies simply present a system that does something, without specifying whether it is intended as a practical device or as a model of cognition. As noted earlier, there may well be some convergence between the two goals. But this does not excuse evading the issue entirely (as some studies do): any proposal for a computational system should state whether its motivation is cognitive, practical, or both. Even if a system is forthrightly declared to be a model of cognition, the question arises, *whose* cognition? Certainly there are significant differences in cognition between the musically trained and untrained—most obviously in performance, but also in perception as well. Undoubtedly, cultural background also plays a role here, though it may be that some aspects of music perception are substantially the same across cultures. (The application of computational modeling techniques to non-Western music has been little explored; see Tzanetakis, Kapur, Schloss, & Wright, 2007, for discussion of some recent work.) Experimental studies are forced to take a stand on this issue, at least implicitly, through their choice of subject populations; in computational research, by contrast, there is a temptation to evade the issue—for example, by referring generically to "the listener."

Notwithstanding these criticisms, the field of computational music cognition can claim some significant accomplishments in its short history. No doubt, progress will be made on the aforementioned challenges as the field continues to mature. The growth of computational music cognition research in the past 40 years has been truly remarkable and seems only to have accelerated in the past decade. It seems likely that this area of research will continue to expand in the coming years, and will become an increasingly important part of the field of music cognition.

Acknowledgment

Thanks are due to Peter Pfordresher for helpful comments on a draft of this chapter.

References

Baroni, M., & Jacobini, R. (1978). *Proposal for a grammar of melody*. Montreal, Canada: Les Presses de l'Université de Montreal.

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

Bigand, E., & Parncutt, R. (1999). Perceiving musical tension in long chord sequences. *Psychological Research*, *62*, 237–254.

Biles, J. A. (1994). GenJam: A genetic algorithm for generating Jazz solos. *Proceedings of the 1994 International Computer Music Conference*. San Francisco, CA: International Computer Music Association.

Bod, R. (2002). A unified model of structural organization in language and music. *Journal of Artificial Intelligence Research*, *17*, 289–308.

Bresin, R. (1998). Artificial neural networks based models for automatic performance of musical scores. *Journal of New Music Research*, *27*, 239–270.

Brooks, F. P., Hopkins, A. L., Neumann, P. G., & Wright, W. V. (1957). An experiment in musical composition. *IRE Transactions on Computers*, EC-6, 175–182.

Bruderer, M. J., McKinney, M. E., & Kohlrausch, A. (2010). The perception of structural boundaries in polyphonic representations of Western popular music. *Musicae Scientiae, Discussion Forum*, *5*, 115–155.

Butler, D. (1989). Describing the perception of tonality in music: A critique of the tonal hierarchy theory and a proposal for a theory of intervallic rivalry. *Music Perception*, *6*, 219–242.

Cambouropoulos, E. (1997). Musical rhythm: A formal model for determining local boundaries, accents and metre in a melodic surface. In M. Leman (Ed.), *Music, gestalt, and computing* (pp. 277–293). Berlin, Germany: Springer-Verlag.

Cambouropoulos, E. (2003). Pitch spelling: A computational model. *Music Perception*, *20*, 411–429.

Cambouropoulos, E. (2006). Musical parallelism and melodic segmentation: A computational approach. *Music Perception*, *23*, 249–267.

Cemgil, A. T., Desain, P., & Kappen, B. (2000). Rhythm quantization for transcription. *Computer Music Journal*, *24/2*, 60–76.

Cemgil, A. T., Kappen, B., Desain, P., & Honing, H. (2000). On Tempo tracking: Tempogram representation and Kalman filtering. *Journal of New Music Research*, *29*, 259–273.

Chai, W. (2004). Melody as a significant musical feature in repertory classification. *Computing in Musicology*, *13*, 51–72.

Chew, E. (2002). The spiral array: An algorithm for determining key boundaries. In C. Anagnostopoulou, M. Ferrand, & A. Smaill (Eds.), *Music and artificial intelligence* (pp. 18–31). Berlin, Germany: Springer.

Chew, E., & Chen, Y.-C. (2005). Real-time pitch spelling using the Spiral Array. *Computer Music Journal*, *29*, 61–76.

Conklin, D., & Anagnostopoulou, C. (2006). Segmental pattern discovery in music. *INFORMS Journal of Computing*, *18*, 285–293.

Conklin, D., & Bergeron, M. (2008). Feature set patterns in music. *Computer Music Journal*, *32/1*, 60–70.

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

Cope, D. (2000). *The algorithmic composer*. Madison, WI: A-R Editions.

Cope, D. (2005). *Computer models of musical creativity*. Cambridge, MA: MIT Press.

Cuddy, L. L. (1997). Tonal relations. In I. Deliege, & J. Sloboda (Eds.), *Perception and cognition of music* (pp. 329–352). London, England: Taylor & Francis.

Cuddy, L. L., & Lunney, C. A. (1995). Expectancies generated by melodic intervals: Perceptual judgments of melodic continuity. *Perception & Psychophysics*, *57*, 451–462.

Dannenberg, R. B., & Derenyi, I. (1998). Combining instrument and performance models for high quality music synthesis. *Journal of New Music Research*, *27*, 211–238.

de Cheveigne, A. (2005). Pitch perception models. In C. J. Plack, A. J. Oxenham, R. R. Fay, & A. N. Popper (Eds.), *Pitch: Neural coding and perception* (pp. 169–233). Berlin, Germany: Springer.

Deliège, I. (2001). Prototype effects in music listening: An empirical approach to the notion of imprint. *Music Perception*, *18*, 371–407.

Desain, P., & Honing, H. (1989). The quantization of musical time: A connectionist approach. *Computer Music Journal*, *13/3*, 56−66.

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

Dixon, S. (2001). Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, *30*, 39−58.

Dixon, S. (2007). Evaluation of the audio beat tracking system BeatRoot. *Journal of New Music Research*, *36*, 39−50.

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

Downie, S. (2006). The Music Information Retrieval Evaluation eXchange (MIREX). *D-Lib Magazine*, *12/12*. (<www.dlib.org/dlib/december06/downie/12downie.html>)

Ebcioglu, K. (1988). An expert system for harmonizing four-part chorales. *Computer Music Journal*, *12/3*, 43−51.

Eck, D. (2001). A positive-evidence model for rhythmical beat induction. *Journal of New Music Research*, *30*, 187−200.

Eerola, T., Järvinen, T., Louhivuori, J., & Toiviainen, P. (2001). Statistical features and perceived similarity of folk melodies. *Music Perception*, *18*, 275−296.

Frankland, B. W., & Cohen, A. J. (2004). Parsing of melody: Quantification and testing of the local grouping rules of "Lerdahl and Jackendoff's A Generative Theory of Tonal Music". *Music Perception*, *21*, 499−543.

Franklin, J. A. (2006). Recurrent neural networks for music computation. *INFORMS Journal on Music Computing*, *18*, 321−338.

Friberg, A., Bresin, R., & Sundberg, J. (2006). Overview of the KTH rule system for muical performance. *Advances in Cognitive Psychology*, *2*, 145−161.

Fujinaga, I. (1998). Machine recognition of timbre using steady state tone of acoustic musical instruments. *Proceedings of the International Computer Music Conference*. San Francisco, CA: International Computer Music Association.

Gabrielsson, A. (1973). Studies in rhythm. *Acta Universitatis Upsaliensis*, *7*, 3−19.

Gasser, M., Eck, D., & Port, R. F. (1999). Meter as mechanism: A neural network that learns metrical patterns. *Connection Science*, *11*, 187−216.

Gjerdingen, R. O. (1994). Apparent motion in music? *Music Perception*, *11*, 335−370.

Goto, M. (2001). An audio-based real-time beat tracking system for music with or without drum-sounds. *Journal of New Music Research*, *30*, 159−171.

Gouyon, F., & Dixon, S. (2005). A review of automatic rhythm description systems. *Computer Music Journal*, *29/1*, 34−54.

Gouyon, F., & Herrera, P. (2003). Determination of the meter of musical audio signals: Seeking recurrences in beat segment descriptors. *Proceedings of the 2003 Audio Engineering Society Convention*. New York, NY: Audio Engineering Society.

Grey, J. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America*, *61*, 1270−1277.

Hamanaka, M., Hirata, K., & Tojo, S. (2006). Implementing 'a generative theory of tonal music.' *Journal of New Music Research*, *35*, 249−277.

Hamanaka, M., & Tojo, S. (2009). Interactive GTTM analyzer. *Proceedings of the International Conference for Music Information Retrieval.* International Society for Music Information Retrieval. (<http://www.ismir.net/>)

Hiller, L. A., & Isaacson, L. M. (1959). *Experimental music*. New York, NY: McGraw-Hill.

Huron, D., & Parncutt, R. (1993). An improved model of tonality perception incorporating pitch salience and echoic memory. *Psychomusicology*, *12*, 154−171.

Jones, M. R., Moynihan, H., MacKenzie, N., & Puente, J. (2002). Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychological Science*, *13*, 313−319.

Juhasz, Z. (2000). Contour analysis of Hungarian folk music in a multidimensional metric-space. *Journal of New Music Research*, *29*, 71−83.

Kashino, K., Nakadai, K., Kinoshita, T., & Tanaka, H. (1998). Application of Bayesian probability networks to musical scene analysis. In D. Rosenthal, & H. Okuno (Eds.), *Computational auditory scene analysis* (pp. 115−137). Mahwah, NJ: Lawrence Erlbaum.

Kassler, M. (1977). Explication of the middleground of Schenker's theory of tonality. *Miscellanea Musicologica*, *9*, 72−81.

Kilian, J., & Hoos, H. (2002). Voice separation: A local optimisation approach. *Proceedings of the International Conference for Music Information Retrieval.* International Society for Music Information Retrieval. (<http://www.ismir.net/>)

Kirlin, P. B., & Utgoff, P. E. (2005). VOISE: Learning to segregate voices in explicit and implicit polyphony. *Proceedings of the International Society for Music Information Retrieval.* International Society for Music Information Retrieval. (<http://www.ismir.net/>)

Kirlin, P. B., & Utgoff, P. E. (2008). An framework for automated Schenkerian analysis. *Proceedings of the International Society for Music Information Retrieval.* International Society for Music Information Retrieval. (<http://www.ismir.net/>)

Klapuri, A. P. (2004). Automatic music transcription as we know it today. *Journal of New Music Research*, *33*, 269−282.

Krumhansl, C. L. (1990). *Cognitive foundations of musical pitch*. New York, NY: Oxford University Press.

Krumhansl, C. L. (1995). Music psychology and music theory: Problems and prospects. *Music Theory Spectrum*, *17*, 53−80.

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, *89*, 334−368.

Lamont, A., & Dibben, N. (2001). Motivic structure and the perception of similarity. *Music Perception*, *18*, 245−274.

Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How people track time varying events. *Psychological Review*, *106*, 119−159.

Large, E. W., & Kolen, J. F. (1994). Resonance and the perception of musical meter. *Connection Science*, *6*, 177−208.

Larson, S. (1997−1998). Musical forces and melodic patterns. *Theory and Practice*, 22−23, 55−71.

Larson, S. (2004). Musical forces and melodic expectations: Comparing computer models and experimental results. *Music Perception*, *21*, 457−498.

Lartillot, O., & St. James, E. (2004). Automating motivic analysis through the application of perceptual rules. *Computing in Musicology*, *13*, 73−91.

Lee, C. (1991). The perception of metrical structure: Experimental evidence and a model. In P. Howell, R. West, & I. Cross (Eds.), *Representing musical structure* (pp. 59−127). London, England: Academic Press.

Leman, M. (1995). *Music and schema theory*. Berlin, Germany: Springer.

Lerdahl, F. (2001). *Tonal pitch space*. Oxford, England: Oxford University Press.

Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge, MA: MIT Press.

Lerdahl, F., & Krumhansl, C. L. (2007). Modeling tonal tension. *Music Perception*, *24*, 329−366.

Longuet−Higgins, H. C., & Steedman, M. J. (1971). On interpreting Bach. *Machine Intelligence*, *6*, 221−241.

Margulis, E. H. (2005). A model of melodic expectation. *Music Perception*, *4*, 663−714.

Marr, D. (1982). *Vision*. New York, NY: Freeman.

Marsden, A. (1992). Modelling the perception of musical voices: A case study in rule-based systems. In A. Marsden, & A. Pople (Eds.), *Computer representations and models in music* (pp. 239−263). London, England: Academic Press.

Marsden, A. (2010). Schenkerian analysis by computer: A proof of concept. *Journal of New Music Research*, *39*, 269−289.

Matsunaga, A., & Abe, J. (2005). Cues for key perception of a melody: Pitch set alone? *Music Perception*, *23*, 153−164.

Mavromatis, P. (2005). A hidden Markov model of melody production in Greek church chant. *Computing in Musicology*, *14*, 93−112.

Mavromatis, P. (2009). Minimum description length modeling of musical structure. *Journal of Mathematics and Music*, *3*, 117−136.

Maxwell, H. J. (1992). An expert system for harmonic analysis of tonal music. In M. Balaban, K. Ebcioglu, & O. Laske (Eds.), *Understanding music with AI* (pp. 335−353). Cambridge, MA: MIT Press.

Mazzola, G., & Zahorka, O. (1994). Tempo curves revisited: Hierarchies of performance fields. *Computer Music Journal*, *18/1*, 40−52.

McAuley, J. D. (1995). *Perception of time as phase: Toward an adaptive-oscillator model of rhythmic pattern processing* (Unpublished doctoral dissertation). Indiana University, Bloomington.

McAuley, J. D., & Semple, P. (1999). The effect of tempo and musical experience on perceived beat. *Australian Journal of Psychology*, *51*, 176−187.

Meredith, D. (2006). The p13 pitch spelling algorithm. *Journal of New Music Research*, *35*, 121−159.

Meredith, D., Lemström, K., & Wiggins, G. (2002). Algorithms for discovering repeated patterns in representations of polyphonic music. *Journal of New Music Research*, *31*, 321−345.

Monahan, C. B., & Carterette, E. C. (1985). Pitch and duration as determinants of musical space. *Music Perception*, *3*, 1−32.

Mozer, M. C. (1994). Neural network music composition by prediction: Exploring the benefits of psychophysical constraints and multiscale processing. *Connection Science*, *6*, 247−280.

Müllensiefen, D., & Frieler, K. (2004). Cognitive adequacy in the measurement of melodic similarity: Algorithmic vs. human judgments. *Computing in Musicology*, *13*, 147−176.

Müller, S., & Mazzola, G. (2003). The extraction of expressive shaping in performance. *Computer Music Journal*, *27/1*, 47−58.

Narmour, E. (1990). *The analysis and cognition of basic melodic structures: The implication-realization model*. Chicago, IL: University of Chicago Press.

Özcan, E., & Erçal, T. (2008). *A genetic algorithm for generating improvised music. In Lecture notes in computer science: Proceedings of the 8th Annual Conference on Artificial Evolution* (pp. 266−277). Berlin, Germany: Springer-Verlag.

Pachet, F. (2002). Interacting with a musical learning system: The continuator. In C. Anagnostopoulou, M. Ferrand, & A. Smaill (Eds.), *Music and artificial intelligence* (pp. 119−132). Berlin, Germany: Springer-Verlag.

Palmer, C., & Krumhansl, K. (1987). Pitch and temporal contributions to musical phrase perception: Effects of harmony, performance timing, and familiarity. *Perception & Psychophysics*, *41*, 505–518.

Palmer, C., & Pfordresher, P. Q. (2003). Incremental planning in sequence production. *Psychological Review*, *110*, 683–712.

Pardo, B., & Birmingham, W. P. (2002). Algorithms for chordal analysis. *Computer Music Journal*, *26/2*, 27–49.

Pardo, B., Shifrin, J., & Birmingham, W. (2004). Name that tune: A pilot study in finding a melody from a sung query. *Journal of the American Society for Information Science and Technology*, *55*, 283–300.

Parncutt, R. (1989). *Harmony: A psychoacoustical approach*. Berlin, Germany: Springer.

Parncutt, R. (1994). A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception*, *11*, 409–464.

Parncutt, R., Sloboda, J. A., Clarke, E., Raekallio, M., & Desain, P. (1997). An ergonomic model of keyboard fingering for melodic fragments. *Music Perception*, *14*, 341–382.

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

Pearce, M., & Wiggins, G. (2006). Expectation in melody: The influence of context and learning. *Music Perception*, *23*, 340–377.

Pfordresher, P. Q., Palmer, C., & Jungers, M. (2007). Speed, accuracy, and serial order in sequence production. *Cognitive Science*, *31*, 63–98.

Pinkerton, R. C. (1956). Information theory and melody. *Scientific American*, *194*, 77–86.

Ponsford, D., Wiggins, G., & Mellish, C. (1999). Statistical learning of harmonic movement. *Journal of New Music Research*, *28*, 150–177.

Povel, D.-J., & Essens, P. (1985). Perception of temporal patterns. *Music Perception*, *2*, 411–440.

Radicioni, D. P., & Lombardo, V. (2005). Computational modeling of chord fingering for string instruments. *Proceedings of the 27th International Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Raphael, C. (2001). A Bayesian network for real-time musical accompaniment. *Neural Information Processing Systems*, *14*, 1433–1439.

Raphael, C. (2002). A hybrid graphical model for rhythmic parsing. *Artificial Intelligence*, *137*, 217–238.

Raphael, C., & Stoddard, J. (2004). Functional harmonic analysis using probabilistic models. *Computer Music Journal*, *28/3*, 45–52.

Rolland, P.-Y. (1999). Discovering patterns in musical sequences. *Journal of New Music Research*, *28*, 334–350.

Rosenthal, D. (1992). Emulation of human rhythm perception. *Computer Music Journal*, *16/1*, 64–76.

Rosenthal, D. F., & Okuno, H. G. (Eds.). (1998). *Computational auditory scene analysis.* Mahwah, NJ: Lawrence Erlbaum Associates.

Rothgeb, J. (1980). Simulating musical skills by digital computer. *Computer Music Journal*, *4/2*, 36–40.

Sadakata, M., Desain, P., & Honing, H. (2006). The Bayesian way to relate rhythm perception and production. *Music Perception*, *23*, 269–288.

Sayegh, S. (1989). Fingering for string instrument with the optimum path paradigm. *Computer Music Journal*, *13*(*3*), 76–84.

Scheirer, E. D. (1998). Tempo and beat analysis of acoustic musical signals. *Journal of the Acoustical Society of America*, *103*, 588–601.

Schellenberg, E. G. (1997). Simplifying the implication-realization model of melodic expectancy. *Music Perception*, *14*, 295−318.

Schenker, H. (1935/1979). *Free composition* (E. Oster, Trans., Ed.). New York, NY: Longman.

Schmuckler, M. (1999). Testing models of melodic contour similarity. *Music Perception*, *16*, 295−326.

Shmulevich, I., & Yli-Harja, O. (2000). Localized key-finding: Algorithms and applications. *Music Perception*, *17*, 65−100.

Smoliar, S. (1980). A computer aid for Schenkerian analysis. *Computer Music Journal*, *4/2*, 41−59.

Steedman, M. (1977). The perception of musical rhythm and meter. *Perception*, *6*, 555−570.

Sundberg, J. (1988). Computer synthesis of music performance. In J. A. Sloboda (Ed.), *Generative processes in music* (pp. 52−69). Oxford, England: Clarendon Press.

Temperley, D. (2001). *The cognition of basic musical structures*. Cambridge, MA: MIT Press.

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

Temperley, D., & Sleator, D. (1999). Modeling meter and harmony: A preference-rule approach. *Computer Music Journal*, *23/1*, 10−27.

Tenney, J., & Polansky, L. (1980). Temporal Gestalt perception in music. *Journal of Music Theory*, *24*, 205−241.

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

Tillmann, B., Bharucha, J. J., & Bigand, E. (2000). Implicit learning of tonality: A self-organizing approach. *Psychological Review*, *107*, 885−913.

Todd, N. P. M. (1989). A computational model of rubato. *Contemporary Music Review*, *3*, 69−88.

Todd, N. P. M. (1992). The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America*, *91*, 3540−3550.

Todd, P. M. (1989). A connectionist approach to algorithmic composition. *Computer Music Journal*, 13/4, 27−43.

Toiviainen, P. (2001). Real-time recognition of improvisations with adaptive oscillators and a recursive Bayesian classifier. *Journal of New Music Research*, *30*, 137−148.

Toiviainen, P., & Krumhansl, C. (2003). Measuring and modeling real-time responses to music: The dynamics of tonality induction. *Perception*, *32*, 741−766.

Typke, R., Giannopoulos, P., Veltkamp, R., Wiering, F., & van Oostrum, R. (2003). Using transportation distances for measuring melodic similarity. *Proceedings of the International Society for Music Information Retrieval.* International Society for Music Information Retrieval. (<http://www.ismir.net/>)

Tzanetakis, G., Kapur, A., Schloss, W. A., & Wright, M. (2007). Computational ethnomusicology. *Journal of Interdisciplinary Music Studies*, *1*, 1−24.

Vos, P. (1999). Key implications of ascending fourth and descending fifth openings. *Psychology of Music*, *27*, 4−17.

Vos, P. G., & Van Geenen, E. W. (1996). A parallel-processing key-finding model. *Music Perception*, *14*, 185−224.

Winograd, T. (1968). Linguistics and the computer analysis of tonal harmony. *Journal of Music Theory*, *12*, 2−49.

Widmer, G. (2002). Machine discoveries: A few simple, robust local expression principles. *Journal of New Music Research*, *31*, 37−50.

Widmer, G., & Tobudic, A. (2003). Playing Mozart by analogy: Learning multi-level timing and dynamics strategies. *Journal of New Music Research*, *32*, 259–268.

Windsor, L. W., & Clarke, E. F. (1997). Expressive timing and dynamics in real and artificial musical performances: Using an algorithm as an analytical tool. *Music Perception*, *15*, 127–152.