Cognitive Science 32 (2008) 418–444

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1080/03640210701864089

A Probabilistic Model of Melody Perception

David Temperley

Eastman School of Music, University of Rochester

Received 15 February 2006; received in revised form 17 April 2007; accepted 19 April 2007

Abstract

This study presents a probabilistic model of melody perception, which infers the key of a melody and also judges the probability of the melody itself. The model uses Bayesian reasoning: For any "surface" pattern and underlying "structure," we can infer the structure maximizing $P(\text{structure} \mid \text{surface})$ based on knowledge of $P(\text{surface, structure})$. The probability of the surface can then be calculated as $\sum P(\text{surface, structure})$, summed over all structures. In this case, the surface is a pattern of notes; the structure is a key. A generative model is proposed, based on three principles: (a) melodies tend to remain within a narrow pitch range; (b) note-to-note intervals within a melody tend to be small; and (c) notes tend to conform to a distribution (or key profile) that depends on the key. The model is tested in three ways. First, it is tested on its ability to identify the keys of a set of folksong melodies. Second, it is tested on a melodic expectation task in which it must judge the probability of different notes occurring given a prior context; these judgments are compared with perception data from a melodic expectation experiment. Finally, the model is tested on its ability to detect incorrect notes in melodies by assigning them lower probabilities than the original versions.

Keywords: Music cognition; Probabilistic modeling; Expectation; Key perception

1. Introduction

In hearing and understanding the notes of a melody, the listener engages in a complex set of perceptual and cognitive processes. The notes must first be identified: the individual partials of the sound must be grouped into complex tones, and these tones must be assigned to the correct pitch categories. The listener then evaluates the notes, judging each one as to whether it is appropriate or probable in the given context. Thus, the listener is able to identify incorrect or deviant notes—whether these are accidental errors by the performer or deliberate surprises injected by the composer. The listener also infers underlying musical structures from the note pattern: the key, the meter, and other kinds of musical information. Finally,

Correspondence should be addressed to David Temperley, Eastman School of Music, 26 Gibbs St., Rochester, NY 14604. E-mail: dtemperley@esm.rochester.edu

the listener forms expectations about what note will occur next and can judge whether these expectations are fulfilled or denied.

All of these processes—note identification, error detection, expectation, and perception of underlying structures—would seem to lend themselves to a probabilistic treatment. The listener is able to judge the probability of different note sequences occurring and brings this knowledge to bear in determining what notes did occur, whether they were intended, and what notes are likely to occur next. The identification of structures such as key and meter could well be viewed from a probabilistic perspective, as well: The listener hears a pattern of notes and must determine the most probable underlying structure (of whatever kind) given those notes.

These cognitive musical processes might be divided into those concerned with the pattern of notes itself, which I will call *surface processes*, and those concerned with the identification of underlying structures, which I will call *structural processes*. Surface processes include pitch identification, error detection, and expectation; structural processes include the perception of meter and key. Notwithstanding this distinction, surface processes and structural processes are closely intertwined. Obviously, identification of underlying structures depends on the identification of the note pattern from which they are inferred. In addition, however, the musical structures that are inferred then guide the perception of the surface. For example, it seems reasonable to suppose—and there is indeed evidence for this, as will be discussed—that our judgment of the key of a melody will affect our expectations of what note will occur next. This raises the possibility that both surface and structural processes might be accommodated within a single cognitive model.

In what follows, I propose a unified probabilistic model of melody perception. The model infers the key of a note pattern; it also judges the probability of the note pattern (and possible continuations of the pattern), thus providing a model of error detection and expectation. (The model only considers the pitch aspect of melody, not rhythm; the rhythmic aspect of melody perception is an enormously complex and largely separate issue, which we will not address here.) The model is designed to simulate the perception of Western tonal music by listeners familiar with this idiom.[1] The model uses the approach of Bayesian probabilistic modeling. Bayesian modeling provides a way of identifying the hidden structures that lie beneath, and give rise to, a surface pattern. At the same time, the Bayesian approach yields a very natural way of evaluating the probability of the surface pattern itself.

I begin by presenting an overview of the model and its theoretical foundation. I then examine, in more detail, the model's handling of three problems: key finding, melodic expectation, and melodic error detection. In each case, I present systematic tests of the model's performance. In the case of key finding, the model's output is compared to "expert" judgments of key on a corpus of folk melodies (and also on a corpus of Bach fugue themes); in the case of expectation, the output is compared to data from a perception experiment (Cuddy & Lunney, 1995). In the case of error detection, the model is tested on its ability to distinguish randomly deformed versions of melodies from the original versions. I will also examine the model's ability to predict scale-degree tendencies and will discuss its relevance to the problem of pitch identification. Finally, I consider some further implications of the model and possible avenues for further development.

2. Theoretical foundation

Bayesian probabilistic modeling has recently been applied to many problems of information processing and cognitive modeling, such as decision-making (Osherson, 1990), vision (Knill & Richards, 1996; Olman & Kersten, 2004), concept learning (Tenenbaum, 1999), learning of causal relations (Sobel, Tenebaum, & Gopnik, 2004), and natural language processing (Eisner, 2002; Jurafsky & Martin, 2000; Manning & Schütze, 2000). To bring out the connections between these domains and the current problem, I present the motivation for the Bayesian approach in a very general way. In many kinds of situations, a perceiver is presented with some kind of surface information (which I will simply call a *surface*) and wants to know the underlying structure or content that gave rise to it (which I will call a *structure*). This problem can be viewed probabilistically, in that a given surface may result from many different structures; the perceiver's goal is to determine the most likely structure, given the surface. Using Bayes' rule, the probability of a structure given a surface can be related to the probability of the surface given the structure:

$$P(\text{structure} \mid \text{surface}) = \frac{P(\text{surface} \mid \text{structure})P(\text{structure})}{P(\text{surface})} \tag{1}$$

The structure maximizing $P(\text{structure} \mid \text{surface})$ will be the one maximizing the expression on the right. Since $P(\text{surface})$, the overall probability of the surface, will be the same for all structures, it can simply be disregarded. To find the most probable structure given a surface, then, we need only know—for all possible structures—the probability of the surface given the structure, and the overall ("prior") probability of the structure:

$$P(\text{structure} \mid \text{surface}) \propto P(\text{surface} \mid \text{structure})P(\text{structure}) \tag{2}$$

By a basic rule of probability, we can rewrite the right-hand side of this expression as the joint probability of the structure and surface:

$$P(\text{structure} \mid \text{surface}) \propto P(\text{surface}, \text{structure}) \tag{3}$$

Also of interest is the overall probability of a surface. This can be formulated as $P(\text{structure}, \text{surface})$, summed over all possible structures:

$$P(\text{surface}) = \sum_{\text{structure}} P(\text{surface}, \text{structure}) \tag{4}$$

To illustrate the Bayesian approach, let us briefly consider two examples in the domain of natural language processing. In speech recognition, the task is to determine the most probable sequence of words given a sequence of phonetic units or "phones"; in this case, then, the sequence of words is the structure and the sequence of phones is the surface. This can be done by estimating, for each possible sequence of words, the prior probability of that word sequence, and the probability of the phone sequence given the word sequence (Jurafsky & Martin, 2000). Another relevant research area has been syntactic parsing; in this case, we can think of the sequence of words as the surface, while the structure is some kind of syntactic representation. Again, to determine the most probable syntactic structure given the words, we can evaluate the probability of different syntactic structures and the probability of the word sequence given

those structures; this is essentially the approach of most recent computational work on syntactic parsing (Manning & Schütze, 2000). Thus, the level of words serves as the structure to the more superficial level of phones and as the surface to the more structural level of syntactic structure.

In the model presented below, the surface is a pattern of notes, while the structure is a key. Much like syntactic parsing and speech recognition, we can use Bayesian reasoning to infer the structure from the surface. We can also use this approach to estimate the probability of the surface itself. As argued earlier, such surface probabilities play an important role in music cognition, contributing to such processes as pitch identification, error detection, and expectation.

As models of cognition, Bayesian models assume that people are sensitive to the frequencies and probabilities of events in their environment. In this respect, the approach connects nicely with other current paradigms in cognitive modeling, such as statistical learning (Saffran, Johnson, Aslin, & Newport, 1999) and statistical models of sentence processing (Juliano & Tanenhaus, 1994; MacDonald, Pearlmutter, & Seidenberg, 1994). The Bayesian perspective also provides a rational basis for the setting of parameters. In calculating *P*(surface, structure) for different structures and surfaces, it makes sense to base these probabilities on actual frequencies of events in the environment. This is the approach that will be taken here.

The application of probabilistic techniques in music research is not new. A number of studies in the 1950s and 1960s applied concepts from information theory—for example, calculating the entropy of musical pieces or corpora by computing transitional probabilities among surface elements (Cohen, 1962; Hiller & Fuller, 1967; Youngblood, 1958). Others have applied probabilistic approaches to the generation of music (Conklin & Witten, 1995; Ponsford, Wiggins, & Mellish, 1999). Very recently, a number of researchers have applied Bayesian approaches to musical problems. Cemgil and colleagues (Cemgil & Kappen, 2003; Cemgil, Kappen, Desain, & Honing, 2000) propose a Bayesian model of meter perception, incorporating probabilistic knowledge about rhythmic patterns and performance timing (see also Raphael, 2002a). Kashino, Nakadai, Kinoshita, and Tanaka (1998) and Raphael (2002b) have proposed Bayesian models of transcription—the process of inferring pitches from an auditory signal. And Bod (2002) models the perception of phrase structure using an approach similar to that of probabilistic context-free grammars. Aside from its general Bayesian approach, the current study has little in common with these earlier studies. No doubt this simply reflects differences between the problems under investigation: Key identification is a very different problem from meter perception, transcription, and phrase perception.[2] It seems clear, however, that these aspects of music perception are not entirely independent, and that a complete model of music cognition will have to integrate them in some way. We will return to this issue at the end of the article.

We now turn to a description of the model. While the model is primarily concerned with perception, it assumes—like most Bayesian models—a generative process as well: we infer a structure from a surface, based on assumptions about how surfaces are generated from structures. Thus, I will begin by sketching the generative model that is assumed.

3. The model

The task of the generative model is to generate a sequence of pitches (no rhythmic information is generated). To develop such a model, we must ask: what kind of pitch sequence makes

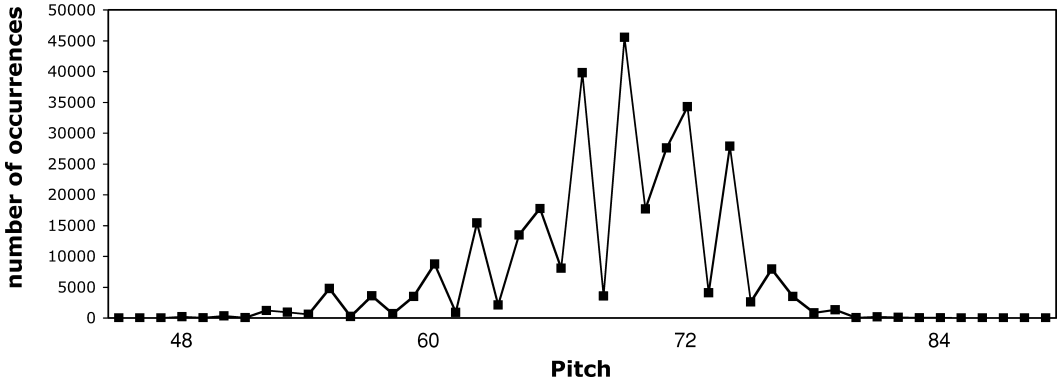

Fig. 1. Distribution of pitches in the *Essen Folksong Collection*. Pitches are represented as integers, with C4 (middle C) = 60.

a likely melody? Perhaps the most basic principle that comes to mind is that a melody tends to be confined to a fairly limited range of pitches. Data were gathered about this from a corpus of 6,149 European folk melodies, the *Essen Folksong Collection* (Schaffrath, 1995). The melodies have been computationally encoded with pitch, rhythm, key, and other information (Huron, 1999).[3]

If we examine the overall distribution of pitches in the corpus (Fig. 1), we find a roughly normal distribution, with the majority of pitches falling in the octave above C4 ("middle C"). Following the usual convention, we will represent pitches as integers, with C4 = 60.) Beyond this general constraint, however, there appears to be an additional constraint on the range of individual melodies. Although the overall variance of pitches in the Essen corpus is 25.0, the variance of pitches within a melody—that is, with respect to the mean pitch of each melody—is 10.6. We can model this situation in a generative way by first choosing a central pitch c for the melody, randomly chosen from a normal distribution, and then creating a second normal distribution centered around c which is used to actually generate the notes. It is important to emphasize that the central pitch of a melody is not the tonal center (the tonic or "home" pitch), but rather the central point of the range. In training, we can estimate the central pitch of a melody simply as the mean pitch rounded to the nearest integer (we assume that c is an integer for reasons that will be explained below). In the Essen collection, the mean of mean pitches is roughly 68 (Ab4), and the variance of mean pitches is 13.2; thus, our normal distribution for choosing c, which we will call the *central pitch profile*, is $N(c; 68, 13.2)$. This normal distribution, like others discussed below, is converted to a discrete distribution taking only integer values. The normal distribution for choosing a series of pitches p_n (the range profile) is then $N(p_n; c, v_r)$. A melody can be constructed as a series of notes generated from this distribution.

A melody generated from a range profile—assuming a central pitch of 68 and variance of 10.6—is shown in Fig. 2a.[4] While this melody is musically deficient in many ways, two problems are particularly apparent. One problem is that the melody contains several wide leaps between pitches. In general, intervals between adjacent notes in a melody are small; this phenomenon of "pitch proximity" has been amply demonstrated as a statistical tendency in

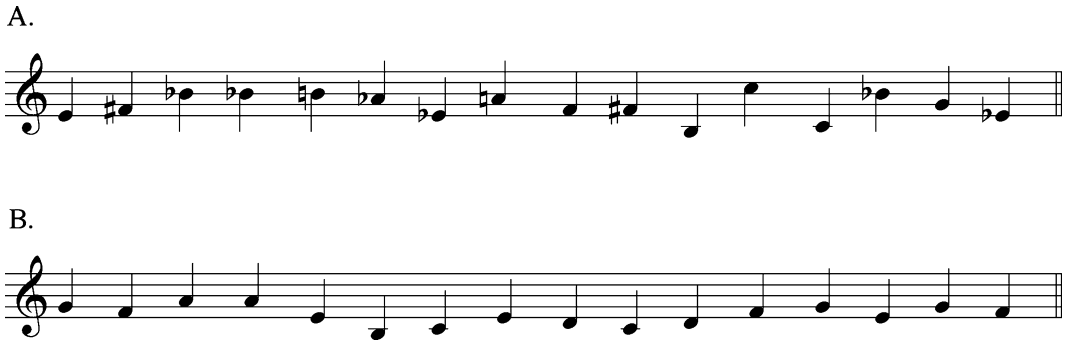

Fig. 2. (A) A melody generated from a range profile. (B) A melody generated from the final model.

actual melodies (von Hippel, 2000; von Hippel & Huron, 2000) and also as an assumption and preference in auditory perception (Deutsch, 1999; Miller & Heise, 1950; Schellenberg, 1996).[5] Figure 3 shows the distribution of "melodic intervals" in the Essen corpus—pitches in relation to the previous pitch; it can be seen that more than half of all intervals are two semitones or less. We can approximate this distribution with a proximity profile—a normal distribution, $N(p_n; p_{n-1}, v_p)$, where p_{n-1} is the previous pitch. We then create a new distribution which is the product of the proximity profile and the range profile. In effect, this "range × proximity" (RP) profile favors melodies which maintain small note-to-note intervals but also remain within a fairly narrow global range. Notice that the RP profile must be recreated at each note, as it depends on the previous pitch. For the first note, there is no previous pitch, so this is generated from the range profile alone.

The range and proximity profiles each have two parameters, the mean and the variance. The mean of the range profile varies from song to song, and the mean of the proximity profile varies from one note to the next. The variances of the two profiles, however—v_r and v_p—do not

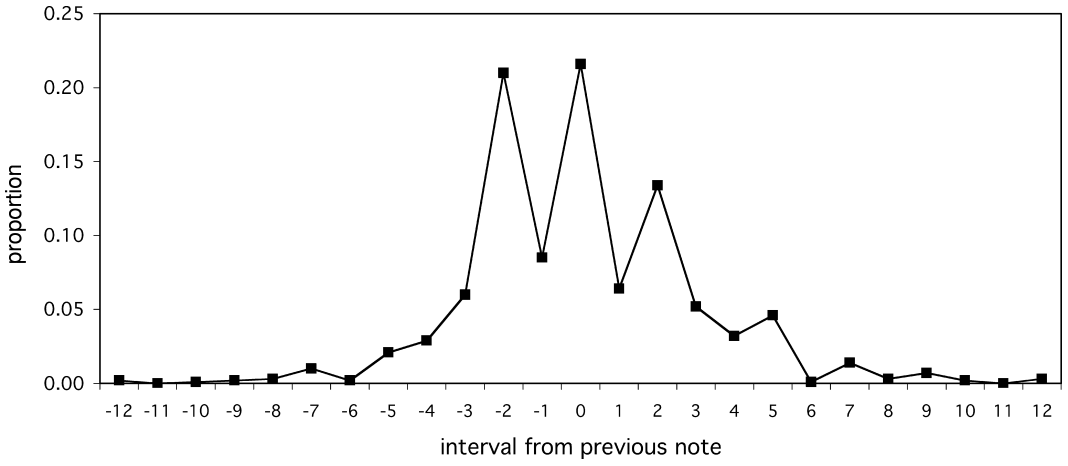

Fig. 3. Melodic intervals in the Essen corpus, showing the frequency of each interval size as a proportion of all intervals. (For example, a value of −2 indicates a note two semitones below the previous note.)

appear to vary greatly across songs; for simplicity, we will assume here that they are constant. The problem is then to estimate them from the Essen data. We could observe the sheer variance of pitches around the mean pitch of each melody, as we did above (yielding a value of 10.6). But this is not the same as v_r; rather, it is affected by both v_p and v_r. (Similarly, the sheer variance of melodic intervals, as shown in Fig. 3, is not the same as v_p.) So another method must be used. It is a known fact that the product of two Gaussians (normal distributions) is another Gaussian, $N(p_n; m_c, v_c)$, whose mean is a convex combination of the means of the Gaussians being multiplied (Petersen & Petersen, 2005):

$$N(p_n; c, v_r)N(p_n; p_{n-1}, v_p) \propto N(p_n; m_c, v_c) \tag{5a}$$

where

$$v_c = v_r v_p/(v_r + v_p) \tag{5b}$$

and

$$m_c = \frac{cv_p + p_{n-1}v_r}{v_r + v_p t} \tag{5c}$$

By hypothesis, the first note of each melody is affected only by the range profile, not the proximity profile. So the variance of the range profile can be estimated as the variance of the first note of each melody around its mean; in the Essen corpus, this yields $v_r = 29.0$. Now consider the case of non-initial notes of a melody where the previous pitch is equal to the central pitch ($p_{n-1} = c$); call this pitch x. (It is because of this step that we need to assume that c is an integer.) At such points, we know from Equation 5c that the mean of the product of the two profiles is also at this pitch:

$$m_c = \frac{xv_p + xv_r}{v_r + v_p} = x \tag{6}$$

Thus, we can estimate v_c as the observed variance of pitches around p_{n-1}, considering only points where $p_{n-1} = c$. The Essen corpus yields a value of $v_c = 5.8$. Now, from Equation 5b, we can calculate v_p as 7.2.

Another improbable aspect of the melody in Fig. 2a is that the pitches do not seem to adhere to any major or minor scale. In a real melody, by contrast (at least in the Western tonal tradition), melodies tend to adhere to the scale of a particular key. A key is a framework of pitch organization, in which pitches are understood to have varying degrees of stability or appropriateness. There are 24 keys: 12 major keys (one named after each pitch class, C, C#, D...B) and 12 minor keys (similarly named). To incorporate key into the model, we adopt the concept of key profiles. A key profile is a 12-valued vector representing the compatibility of each pitch class with a key (Krumhansl, 1990; Krumhansl & Kessler, 1982). In the current model, key profiles are construed probabilistically: the key profile values represent the probability of a pitch class occurring, given a key. The key profile values were set using the Essen corpus; the corpus provides key labels for each melody, allowing pitch class distributions to be tallied in songs of each key. This data were then aggregated over all major keys and all minor keys, producing data as to the frequency of scale degrees, or pitch classes in relation to a key. (For example, in C major, C is scale degree 1, C# is #1, and D is 2;

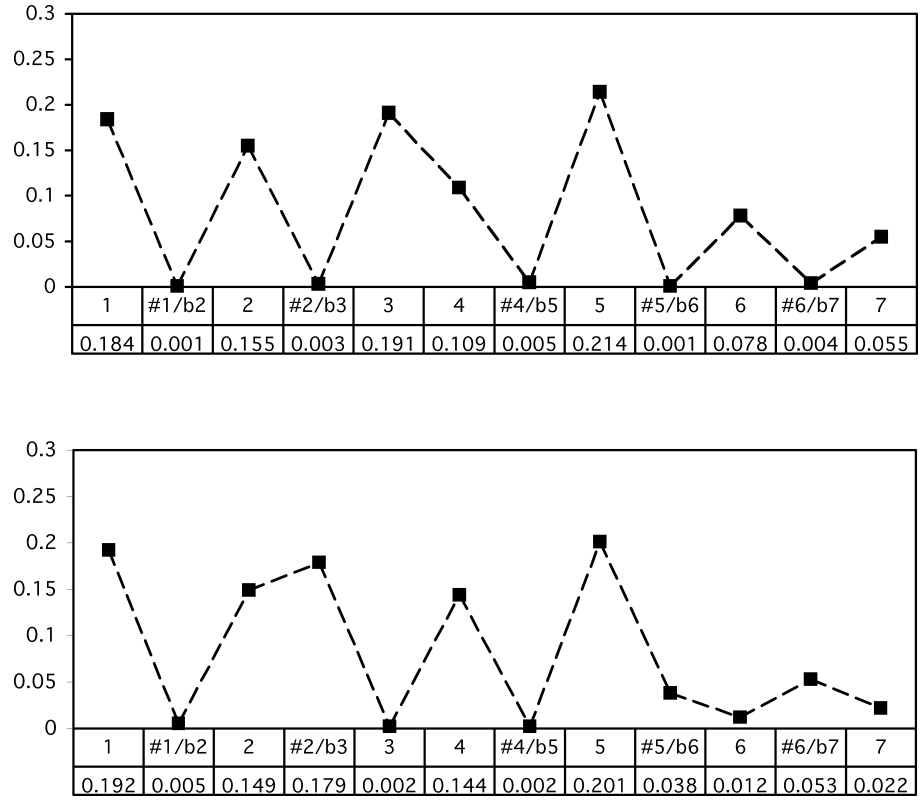

Fig. 4. Key profiles generated from the *Essen Folksong Collection* for major keys (above) and minor keys (below).

in C# major, C# is 1; and so on.) The resulting key profiles are shown in Fig. 4. The profiles show that, for example, 18.4% of notes in major-key melodies are scale degree 1. The profiles reflect conventional musical wisdom, in that pitches belonging to the major or minor scale of the key have higher values than other pitches, and pitches of the tonic chord (the 1, 3, and 5 degrees in major or the 1, b3, and 5 degrees in minor) have higher values than other scalar ones.

The key profiles in Fig. 4 can be used to capture the fact that the probability of pitches occurring in a melody depends on their relationship to the key. However, key profiles only represent pitch class, not pitch: they do not distinguish between middle C, the C an octave below, and the C an octave above. We address this problem by duplicating the key profiles over several octaves. We then multiply the key profile distribution by the RP distribution, normalizing the resulting combined distribution so that the sum of all values is still 1; we will call this the *RPK profile*. Fig. 5 shows an RPK profile, assuming a key of C major, a central pitch of 68 (Ab4), and a previous note of C4. In generating a melody, then, we must construct the RPK profiles anew at each point, depending on the previous pitch. (For the first note, we simply use the product of the range and key profiles.) Fig. 2b shows a melody generated by this method, assuming a key of C major and a central pitch of Ab4. It can be seen that the pitches are all within the C major scale, and that the large leaps found in Fig. 2a are no longer present.

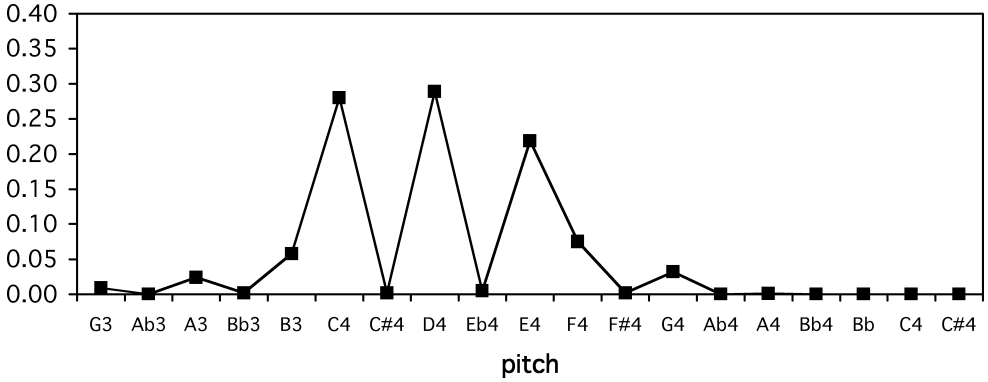

Fig. 5. An RPK profile, assuming a central pitch of Ab4, a previous pitch of C4, and a key of C major.

The generative process thus requires the choice of a key and central pitch and the generation of a series of pitches. The probability of a pitch occurring at any point is given by its RPK profile value: the normalized product of its range profile value (given the central pitch), its proximity profile value (given the previous pitch), and its key profile value (given the chosen key).[6] The model can be represented graphically as shown in Fig. 6. The joint probability of a pitch sequence with a key k and a central pitch c is

$$P(\text{pitch sequence}, k, c) = P(k)P(c)\prod_n P(p_n \mid p_{n-1}, k, c) = P(k)P(c)\prod RPK_n \qquad (7)$$

where p_n is the pitch of the nth note and RPK_n is its RPK profile value. As noted earlier, $P(c)$ is determined by the central pitch profile. (In principle, c could take an infinite range of integer values; but when c is far removed from the pitches of the melody, its joint probability with the melody is effectively zero.) As for $P(k)$, we assume that all keys are equal in prior probability, since most listeners—lacking "absolute pitch"—are incapable of identifying keys in absolute terms; however, we assign major keys a higher probability than minor keys, reflecting the higher proportion of major-key melodies in the Essen collection. ($P(k) = .88/12$ for each major key, .12/12 for each minor key.)

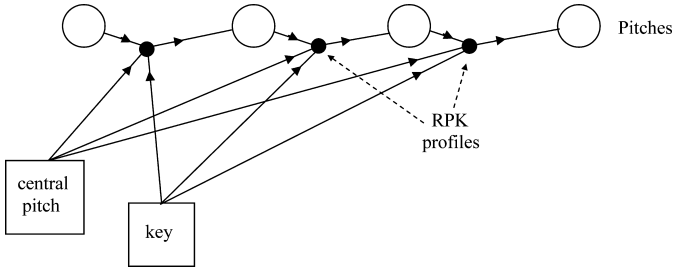

Fig. 6. A graphical representation of the model. The central pitch, key, and pitches are random variables; the RPK profiles are deterministically generated from the key, central pitch, and previous pitch.

The joint probability of a pitch sequence with a key (which will be important in what follows) sums the quantity in Equation 7 over all central pitches:

$$P(\text{pitch sequence}, k) = \sum_c \left[P(k)P(c) \prod\nolimits_n RPK_n \right]$$
$$= P(k) \sum_c \left[P(c) \prod\nolimits_n RPK_n \right] \quad (8)$$

Finally, the overall probability of a melody sums the quantity in Equation 7 over all central pitches and keys:

$$P(\text{pitch sequence}) = \sum_{k,c} \left[P(k)P(c) \prod\nolimits_n RPK_n \right] \quad (9)$$

Essentially, the model has five parameters: The mean of the central pitch profile; the variances of the central pitch profile, range profile, and proximity profile; and the probability of a major key versus a minor key.[7] The variances of the range and proximity profiles determine the "weight" of these factors in the RPK profile. If the proximity variance is very high, pitch proximity will have little effect on the RPK profile and there will be little pressure for small melodic intervals; if the range variance is very high, range will have little effect. If both the range and proximity variances are large, neither range nor pitch proximity will have much weight and the RPK profile will be determined almost entirely by the key profile.

The parameter values proposed above were extracted directly from the Essen corpus. Another approach to parameter setting is also possible, using the technique of maximum likelihood estimation (MLE). Since the model assigns a probability to any melody it is given (Equation 9), one might define the optimal parameters as those which assign highest probability to the data. Using a random sample of 10 melodies from the Essen corpus, a simple optimization approach was used to find the MLE values for the parameters. Starting with random initial values, one parameter was set to a wide range of different values, and the value yielding the highest probability for the data was added to the parameter set; this was done for all five parameters, and the process was iterated until no further improvement was obtained.[8] The entire process was repeated five times with different initial values; all five runs converged to the same parameter set, shown in Table 1. This process is only guaranteed to find a local optimum, not a global optimum, but the fact that all five runs converged on the same parameter set suggests that this is indeed the global optimum. The optimized parameter set assigns a log probability to the 10-song training set of −964.5, whereas the original parameter set assigns a log probability of −976.2. Thus, the optimized parameter set achieves a slightly higher probability, though the difference is very small (1.2%). (By contrast, the five sets of random values used to initialize the optimization yielded an average log probability of −1553.7.)

Having presented the generative model, we now examine how it might be used to model three perceptual processes: key identification, melodic expectation, and error detection.

Table 1
Parameter values for three versions of the model

Parameter	Value Estimated From Essen Corpus	Value Optimized on 10-Song Training Set	Value Optimized on Cuddy and Lunney (1995) Data
Central pitch mean	68	68	64
Central pitch variance	13.2	5.0	13.0
Range variance	29.0	23.0	17.0
Proximity variance	7.2	10.0	70.0
Probability of a major key	0.88	0.86	0.66
Last note factor (on last note, degree 1 in key profile is multiplied by this value)	—	—	20.0

4. Testing the model on key finding

The perception of key has been the focus of a large amount of research. Experimental studies have shown, first of all, that listeners—both musically trained and untrained—are sensitive to key and that there is a good deal of agreement in the way key is perceived (Brown, Butler, & Jones, 1994; Cuddy, 1997; Krumhansl, 1990). Other research has focused on the problem of how listeners infer a key from a pattern of notes—sometimes called the "key finding" problem; a number of models of this process have been put forth, both in psychology and in artificial intelligence (see Temperley, 2001, for a review). We will just consider two well-known models here and will compare their performance to that of the current probabilistic model.

Longuet-Higgins and Steedman (1971) proposed a model for determining the key of a monophonic piece. Longuet-Higgins and Steedman's model is based on the conventional association between keys and scales. The model proceeds left to right from the beginning of the melody; at each note, it eliminates all keys whose scales do not contain that note. When only one key remains, that is the chosen key. If the model gets to the end of the melody with more than one key remaining, it looks at the first note and chooses the key of which that note is scale degree 1 (or, failing that, scale degree 5). If at any point all keys have been eliminated, the "first note" rule again applies. An alternative approach to key finding was proposed by Krumhansl and Schmuckler (described most fully in Krumhansl, 1990). The Krumhansl-Schmuckler key-finding algorithm is based on a set of key profiles representing the compatibility of each pitch class with each key. (The key profiles were derived from experiments by Krumhansl & Kessler, 1982, in which listeners heard a context establishing a key followed by a single pitch and judged how well the pitch "fit" given the context.) The key profiles are shown in Fig. 7; as before, pitch classes are identified in relative or "scale degree" terms. (Note the very strong qualitative similarity between the Krumhansl-Kessler profiles and those derived from the Essen collection, shown in Fig. 4.) Given these profiles, the Krumhansl-Schmuckler algorithm judges the key of a piece by generating an input vector for the piece; this is, again, a vector of 12 values, showing the total duration of each pitch class in the piece. The correlation is then calculated between each key profile vector and the input vector; the key whose profile yields the highest correlation value is the preferred key.

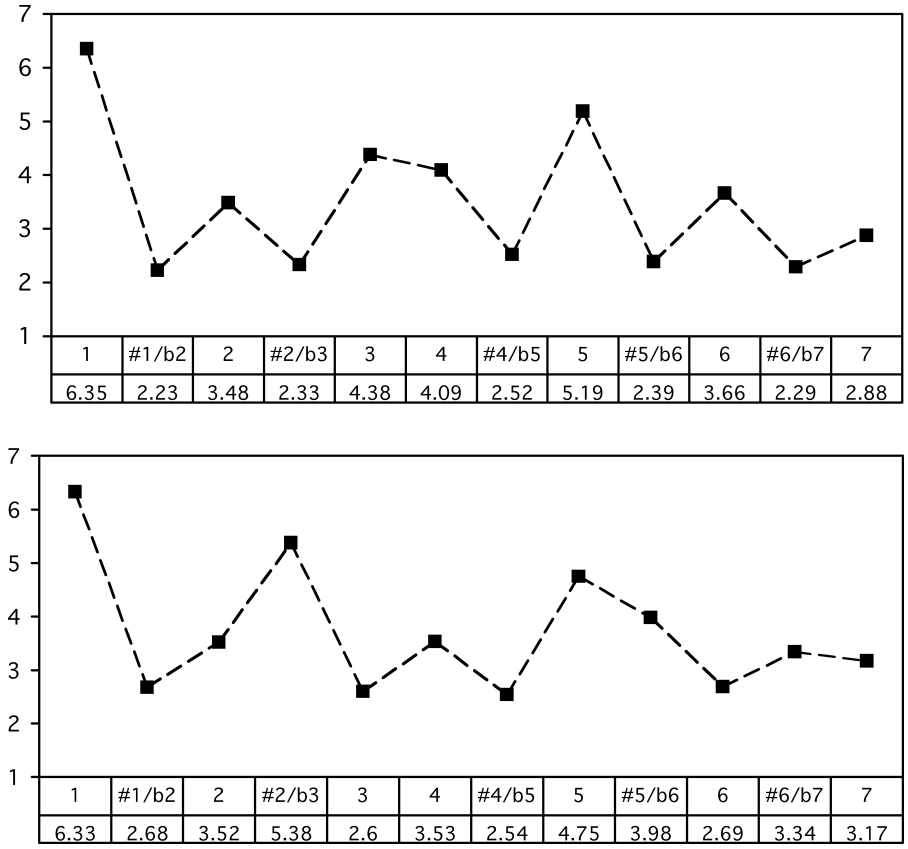

Fig. 7. Key profiles from Krumhansl and Kessler (1982) for major keys (above) and minor keys (below).

We now consider how the probabilistic model proposed above could be used for key finding. The model's task, in this case, is to judge the most probable key given a pitch sequence. It can be seen from Equations 3 and 8 that, for a given key k_x,

$$P(k_x \mid \text{pitch sequence}) \propto P(\text{pitch sequence}, k_x) = P(k_x) \sum_c \left[P(c) \prod\nolimits_n RPK_n \right] \qquad (10)$$

The most probable key given a melody is the one maximizing this expression.[9]

The model was tested on two different corpora. First, it was tested using the *Essen Folksong Collection*—the same corpus described earlier and used for setting the model's parameters. A 65-song test set was extracted from the corpus (this portion of the corpus was not used in parameter setting).[10] The task was simply to judge the key of each melody. The model judged the key correctly for 57 of the 65 melodies (87.7%; see Table 2). The same corpus was then used to test the Longuet-Higgins/Steedman and Krumhansl-Schmuckler models (using my own implementations). The Longuet-Higgins/Steedman model identified the correct key on 46 out of 65 melodies, or 70.8% correct; the Krumhansl-Schmuckler model identified the correct key on 49 out of 65, or 75.4% correct. The second test used a corpus that has been widely

Table 2
Results of key-finding tests of the current model ("probabilistic model") and other models on two different corpora

Test Corpus and Model	# Correct	% Correct
65-Song Essen folksong test set		
Longuet-Higgins/Steedman model	46	70.8
Krumhansl-Schmuckler model	49	75.4
Probabilistic model	57	87.7
48 Fugue subjects from Bach's *Well-Tempered Clavier*		
Longuet-Higgins/Steedman model	48	100.0
Krumhansl-Schmuckler model	32	66.7
Vos and Van Geenen (1996) model	39	81.2
Temperley (2001) model	43	89.6
Probabilistic model	40	83.3
Probabilistic model with adjusted parameters	44	91.7

used for testing in other key finding studies—the 48 fugue subjects of Bach's *Well-Tempered Clavier* ("subject" in this case means the theme of a fugue). This corpus was first used by Longuet-Higgins and Steedman, whose model chose the correct key in all 48 cases (100.0% correct). Results for the current model and four other models are shown in Table 2.[11] The current model chose the correct main key in 40 of the 48 cases (83.3% correct). Inspection of the results suggested that some of the model's errors were due to a problem with the key profiles: in minor keys, the b7 degree has a higher value than 7, whereas in the Bach corpus (as in classical music generally), 7 is much more commonly used in minor keys than b7. When scale degree b7 was given a value of .015 in the minor profile and scale degree 7 was given .060, and the preference for major keys was removed, the correct rate of the model was increased to 44 out of 48 cases (91.7% correct).

Altogether, the model's key-finding performance seems promising. It is probably impossible for a purely "distributional" key-finding model of any kind to achieve perfect performance; in some cases, the temporal arrangement of pitches must also be considered (see Temperley, 2004, for further discussion of this issue). One might wonder, also, whether the "expert" key judgments in the Essen collection and the Bach fugues would always correspond to those of human listeners. While the general correspondence between listener judgments and expert judgments with regard to key has been established (Cuddy, 1997), they might not necessarily coincide in every case. This concern will be addressed in the next section, where we compare the model's judgments with experimental perception data.

5. Testing the model on expectation and error detection

As well as modeling the analytical process of key finding, it was suggested earlier that a probabilistic model of melody could shed interesting light on surface processes of note identification and interpretation. In key finding, the model found the structure maximizing P(surface, structure); using Equation 3, we took this to indicate the most probable structure

given the surface. Now, we use the same quantity, but summed over all possible structures, indicating the probability of the surface itself—that is, the probability of a pitch sequence. I will argue here that the probability of a pitch sequence, defined in this way, is a concept with explanatory relevance to a variety of musical phenomena.

One very important aspect of melody perception is expectation. It is well known that in listening to a melody, listeners form expectations as to what note is coming next; the creation, fulfillment, and denial of such expectations has long been thought to be an important part of musical affect and meaning (Meyer, 1956; Narmour, 1990). Melodic expectation has been the subject of a large amount of psychological research. As noted at the outset of this study, expectation could well be considered a fundamentally probabilistic phenomenon: A judgment of the "expectedness" of a note could be seen as an estimate of its probability of occurring in that context. While this point has been observed before—for example, Schellenberg, Adachi, Purdy, and McKinnon (2002) define expectation as "anticipation of an event based on its probability of occurring" (p. 511)—no attempt has yet been made to model melodic expectation in probabilistic terms. With regard to experimental research, most studies have used one of two paradigms: a perception paradigm, in which subjects are played musical contexts followed by a continuation tone and are asked to judge the expectedness of the tone (Cuddy & Lunney, 1995; Krumhansl, Louhivuori, Toiviainen, Järvinen, & Eerola, 1999; Schellenberg, 1996; Schmuckler, 1989); and a production paradigm, in which listeners are given a context and asked to produce the tone (or series of tones) that they consider most likely to follow (Carlsen, 1981; Lake, 1987; Larson, 2004; Povel, 1996; Thompson, Cuddy, & Plaus, 1997; Unyk & Carlsen, 1987). For our purposes, perception data seem most valuable, since they indicate the relative expectedness of different possible continuations, whereas production data only indicate continuations that subjects judged as most expected. Of particular interest are data from a study by Cuddy and Lunney (1995). In this study, subjects were played a context of two notes played in sequence (the implicative interval), followed by a third note (the continuation tone) and were asked to judge the third note given the first two on a scale of 1 (*extremely bad continuation*) to 7 (*extremely good continuation*). Eight different contexts were used: ascending and descending major second, ascending and descending minor third, ascending and descending major sixth, and ascending and descending minor seventh (see Fig. 8). Each two-note context was followed by 25 different continuation tones, representing all tones within an octave above or below the second tone of the context (which was always either C4 or F#4). For each condition (context plus continuation tone), Cuddy and Lunney reported the average rating, thus yielding 200 data points in all. These data will be considered further below.

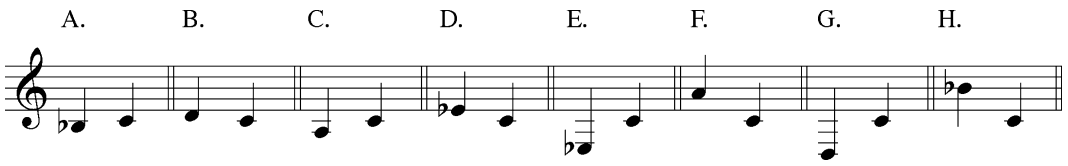

Fig. 8. Two-note contexts used in Cuddy and Lunney (1995). (A) Ascending major second, (B) descending major second, (C) ascending minor third, (D) descending minor third, (E) ascending major sixth, (F) descending major sixth, (G) ascending minor seventh, (H) descending minor seventh. The continuation tone could be any tone within one octave above or below the second context tone.

A number of models of expectation have been proposed and tested on experimental perception data (Cuddy & Lunney, 1995; Krumhansl et al., 1999; Schellenberg, 1996, 1997; Schmuckler, 1989). The usual technique is to use multiple regression. Given a context, each possible continuation is assigned a score that is a linear combination of several variables; multiple regression is used to fit these variables to experimental judgments in the optimal way. Schmuckler (1989) played excerpts from a Schumann song followed by various possible continuations (playing melody and accompaniment separately and then both together); regarding the melody, subjects' judgments correlated with Krumhansl and Kessler's (1982) key profiles and with principles of melodic shape proposed by Meyer (1973). Other work has built on the Implication-Realization theory of Narmour (1990), which predicts expectations as a function of the shape of a melody. Narmour's theory was quantified by Krumhansl (1995) and Schellenberg (1996) to include five factors: registral direction, intervallic difference, registral return, proximity, and closure (these factors will not be explained in detail here). Schellenberg (1996) applied this model to experimental data in which listeners judged possible continuations of excerpts from folk melodies. Cuddy and Lunney (1995) modeled their expectation data (described above) with these five factors; they also included predictors for pitch height, tonal strength (the degree to which the pattern strongly implied a key—quantified using Krumhansl & Kessler's key profile values), and tonal region (the ability of the final tone to serve as a tonic, given the two context tones). On Cuddy and Lunney's experimental data, this model achieved a correlation of .80. Schellenberg (1997) found that a simpler version of Narmour's theory achieved equal or better fit to expectation data than the earlier five-factor version. Schellenberg's simpler model consists of only two factors relating to melodic shape—a "proximity" factor, in which pitches close to previous pitches are more likely, and a "reversal" factor, which favors a change of direction after large intervals—as well as the predictors of pitch height, tonal strength, and tonal region used by Cuddy and Lunney. Using this simplified model, Schellenberg reanalyzed Cuddy and Lunney's data and found a correlation of .851. Whether the five-factor version of Narmour's model or the simplified two-factor version provides a better fit to experimental data has been a matter of some debate (Krumhansl et al., 1999; Schellenberg et al., 2002).

To test the current model against Cuddy and Lunney's (1995) data, we must reinterpret that data in probabilistic terms. There are various ways that this might be done. One could interpret subjects' ratings as probabilities (or proportional to probabilities) of different continuations given a previous context; one could also interpret the ratings as logarithms of probabilities or as some other function of probabilities. There seems little a priori basis for deciding this issue. Initially, ratings were treated as directly proportional to probabilities, but this yielded poor results; treating the ratings as logarithms of probabilities gave much better results, and we adopt that approach in what follows. Specifically, each rating is taken to indicate the log probability of the continuation tone given the previous two-note context. Under the current model, the probability of a pitch p_n given a previous context $(p_o \ldots p_{n-1})$ can be expressed as

$$P(p_n \mid p_o \ldots p_{n-1}) = P(p_o \ldots p_n)/P(p_o \ldots p_{n-1}) \tag{11}$$

where $P(p_o \ldots p_n)$ is the overall probability of the context plus the continuation tone, and $P(p_o \ldots p_{n-1})$ is the probability of just the context. An expression indicating the probability of a sequence of tones was given in Equation 9; this can be used here to calculate both

$P(p_o \ldots p_{n-1})$ and $P(p_o \ldots p_n)$. For example, given a context of (Bb4, C4) and a continuation tone of D4, the model's expectation judgment would be $\log[P(\text{Bb4, C4, D4})/P(\text{Bb4, C4})] = -1.973$.

The model was run on the 200 test items in Cuddy and Lunney's data, and its outputs were compared with the experimental ratings for each item.[12] Using the optimized parameters gathered from the Essen corpus, the model yielded the correlation $r = 0.744$. It seemed reasonable, however, to adjust the parameters to achieve a better fit to the data. This is analogous to what is done in a multiple regression—as used by Cuddy and Lunney (1995), Schellenberg (1997), and others—in which the weight of each predictor is set to optimally fit the data. It was apparent from the experimental data, also, that many highly rated patterns were ones in which the final tone could be interpreted as the tonic of the key. (This trend was also noted by Cuddy & Lunney and Schellenberg, who introduced a special tonal region factor to account for it.) This factor was incorporated into the current model by using special key profiles for the continuation tone, in which the value for the tonic pitch is much higher than usual. This parameter was added to the original five parameters, and all six parameters were then fit to Cuddy and Lunney's data using the same optimization method described in section 3 (see Table 1).[13] With these adjustments, the model achieved a score of $r = .883$, better than both Cuddy and Lunney's model (.80) and Schellenberg's (.851). Figure 9 shows Cuddy and Lunney's data along with the model's output, using the optimized parameters, for two of their eight context intervals (ascending major second and descending major sixth).

One interesting emergent feature of the current model is its handling of "post-skip reversal" or "gap fill." It is a well-established musical principle that large leaps in melodies tend to be followed by a change of direction. Some models incorporate post-skip reversal as an explicit preference: it is reflected, for example, in the registral direction factor of Narmour's model and in the reversal factor of Schellenberg's two-factor model. However, von Hippel and Huron (2000) have suggested that post-skip reversal might simply be an artifact of "regression to the mean." A large interval is likely to take a melody close to the edge of its range; the preference to stay close to the center of the range will thus exert pressure for a change of direction. The current model follows this approach. While there is no explicit preference for post-skip reversal, a context consisting of a large descending interval like A4–C4 is generated with highest probability by a range centered somewhat above the second pitch; given such a range, the pitch following C4 is most likely to move closer to the center, thus causing a change in direction. The preference for ascending intervals following a descending major sixth, though slight, can be seen in Fig. 9 in Cuddy and Lunney's data as well as in the model's predictions. (Values for ascending—positive—intervals are somewhat higher than for descending ones.) It appears that such an indirect treatment of post-skip reversal as an artifact of range and proximity constraints can model expectation data quite successfully.

The influence of key on the model's behavior is also interesting to consider. For example, given the ascending-major-second context (Fig. 9), compare the model's judgments (and the experimental data) for continuations of a descending major second (-2) and descending minor second (-1). Proximity would favor -1, and range would seem to express little preference. So why does the model reflect a much higher value for -2? The reason surely lies in the influence of key. Note that the model does not make a single, determinate key judgment here, nor should it. A context such as Bb4–C4 is quite ambiguous with regard to key; it might imply Bb major,

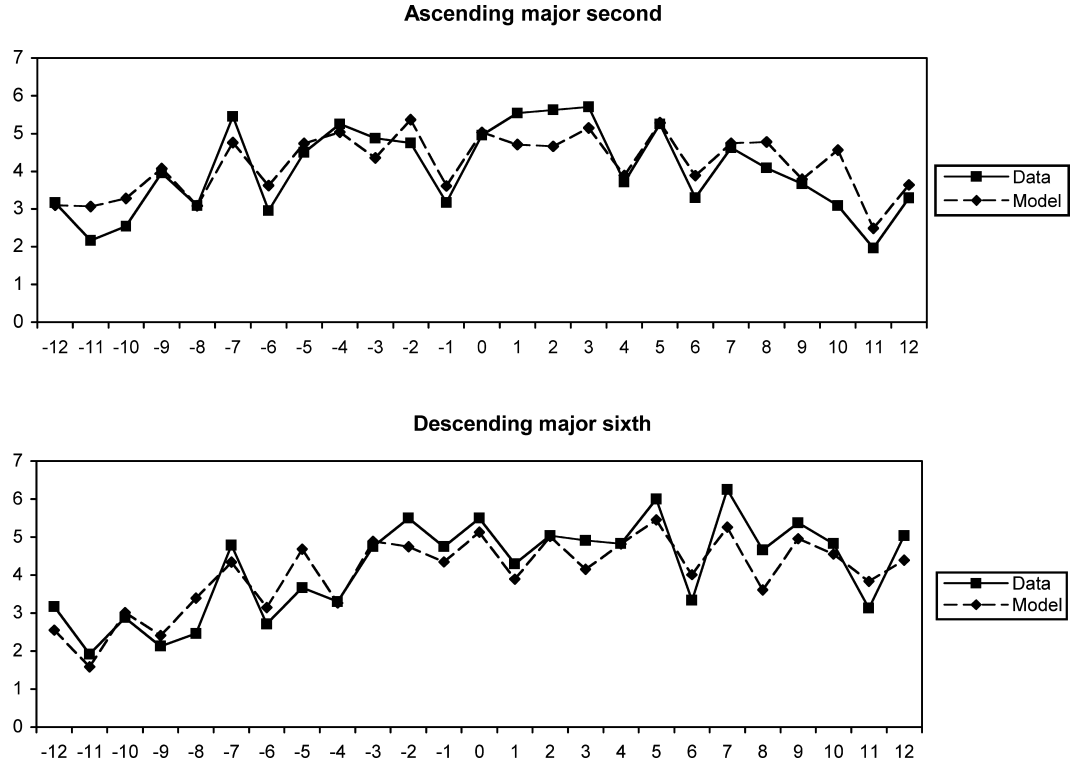

Fig. 9. Expectation data from Cuddy and Lunney (1995) and the model's predictions. Data are shown for two two-tone contexts, ascending major second (Bb3–C4) and descending major sixth (A4–C4). The horizontal axis indicates continuation tones in relation to the second context tone. The vertical axis represents mean judgments of expectedness for the continuation tone given the context, from Cuddy and Lunney's experimental data and as predicted by the model. (The model's output here has been put through a linear function which does not affect the correlation results, but allows easier comparison with the experimental data.)

Bb minor, Eb major, G minor, or other keys. In each of these cases, however, a continuation of −2 (moving back to Bb4) remains within the scale of the key, whereas a continuation of −1 (moving to B4) does not. Thus, key plays an important role in the model's expectation behavior, even when the actual key is in fact quite ambiguous. The fact that the experimental data also reflects a higher rating for −2 than for −1 suggests that this is the case perceptually as well.

We can further understand the model's expectation behavior by examining its handling of scale degree tendencies. It is well known that certain degrees of the scale have tendencies to move to other degrees (Aldwell & Schachter, 2003); such tendencies have been shown to play an important role in melodic expectation (Huron, 2006; Larson, 2004; Lerdahl, 2001). We can represent the tendency of a degree SD1 in terms of the degree SD2 that is most likely to follow (we call this the "primary follower" of SD1) along with its probability of following (the "tendency" value of SD1). (We do not allow a degree to be its own primary follower; tones often do repeat, but this is not usually considered a case of melodic "motion.") To model this, we use a version of the model with a very high range variance, thus minimizing the effect of the range profile; this seems appropriate, since the inherent tendency of a scale degree presumably

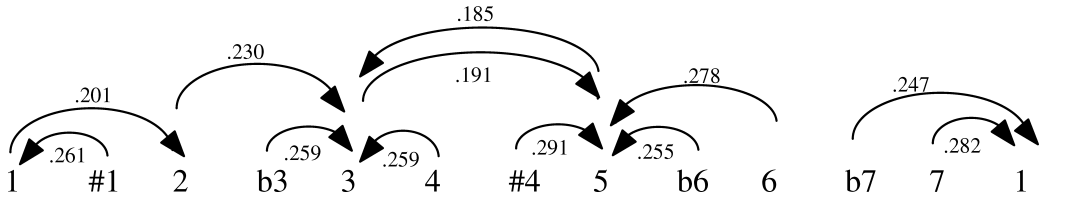

Fig. 10. Scale-degree tendencies as predicted by the model. Arrows indicate, for each scale degree, the "primary follower"—the scale degree that is most likely to follow; the number on the arrow indicates the "tendency value"—the primary follower's probability of following.

depends only on the scale degree itself and should not be affected by any larger context. In effect, then, scale degree tendencies are determined by pitch proximity and by the overall probability of each degree in the scale (as represented in the key profiles). Melodies were created consisting of a one-octave C major scale repeated three times, to establish a strong key context, followed by all possible pairs of pitches (SD1, SD2) within the octave 60–72, representing all pairs of scale degrees. For each SD1, the tendency value was defined as the maximal value of $P(\text{SD2} \mid \text{SD1})$, and the primary follower was defined as the SD2 yielding this maximal value. Figure 10 shows the primary follower and tendency value for each scale degree.

For the most part, the results in Fig. 10 accord well with the usual assumptions of music theory. For degrees outside the scale (#1, b3, #4, b6, and b7), the primary follower is an adjacent degree of the major scale: #1 resolves to 1, #4 resolves to 5, and so on. (The exception is b7, whose primary follower is 1.) These chromatic degrees also have relatively high tendency values, above the average of .245, reflecting the strong expectation for these tones to resolve in a specific way. Turning to the scalar degrees (the degrees of the major scale), it can be seen that they tend toward adjacent scalar degrees, except for 5 (which tends toward 3) and 3 (which tends toward 5). We find relatively high tendency values for 4 (.259) and 7 (.282); both of these degrees are strongly inclined to resolve to a particular scale tone (they are sometimes called *tendency tones*), no doubt due to the fact that each one is a half step from a note of the tonic triad. The lowest tendency values are for 1, 3, and 5, the three degrees of the tonic triad. Roughly speaking, the tendency values are the inverse of the key profile values, with tonic-triad degrees having the lowest values, other scalar degrees having higher values, and chromatic degrees having the highest values. One reason that degrees with lower probability have higher tendency values is that the probability of staying on the same note for such degrees is much smaller, leaving more probability mass for motion to other degrees. (Put simply: one reason why a chromatic note seems to want to move is the simple fact that it is unlikely to stay in the same place.) On the whole, then, the conventional tendencies of scale degrees can be predicted quite well, simply as a function of pitch proximity and the overall probability of different degrees. It would be interesting to compare these predictions with empirical measures of scale-degree tendency, but this will not be undertaken here.[14]

Another kind of phenomenon that is illuminated by the current model could be broadly described as *note error detection*. It seems uncontroversial that most human listeners have some ability to detect errors—"wrong notes"—even in an unfamiliar melody. This ability has been shown in studies of music performance; in sight-reading an unfamiliar score, performers often unconsciously correct anomalous notes (Sloboda, 1976). The ability to detect errors

presumably depends on knowledge of the probabilities of different notes occurring; and it seems plausible that the principles of melodic construction discussed above—principles of key, range, and pitch proximity—are brought to bear in this process.

The model's ability to detect errors was tested using the 65-song Essen folksong sample described above. The model was given the original Essen melodies as well as randomly distorted versions of the same melodies, produced by randomly choosing one note and replacing it by a random pitch chosen from a uniform distribution over the range of the melody (between the lowest and highest pitch). The question was whether the model could reliably assign a higher probability to the correct versions as opposed to the distorted versions.[15] Each of the 65 melodies in the sample was randomly distorted; to ensure a statistically reliable sample, the process was repeated 10 times, yielding a total of 650 trials. In each trial, the model's analyses for the correct version and the distorted version were compared simply with regard to the total probability given to the melody (as defined in Equation 9) to see which version was assigned higher probability. In effect, then, the model simply judged which of a pair of melodies was more likely to contain an error, without expressing any opinion as to exactly where the error was. (In this test, the parameter values optimized on the Essen corpus were used.) The model assigned the correct version of the melody higher probability than the distorted version in 596 out of 650 trials (91.7%). This level of performance seems promising. Probably, not all random errors of this type would be identifiable as errors even by humans; whether the model's ability is comparable to that of human listeners remains to be tested.

A final aspect of melody perception deserving mention here is the actual identification of notes. The extraction of note information from an auditory signal, sometimes known as *music recognition* or *transcription*, is a complex process, involving the grouping of partials (individual frequencies) into complex tones and correct categorization of these complex tones into pitch categories. A number of models of this process have been proposed, both for monophonic input and for the much more difficult problem of polyphonic input (Bello, Monti, & Sandler, 2000; Godsmark & Brown, 1999; Kashino et al., 1998; Martin, 1996; Raphael, 2002b). It seems likely that a model such as the one proposed above could contribute to this task by evaluating the probability of different note patterns. While several Bayesian models of transcription have been proposed (notably Kashino et al., 1998, and Raphael, 2002b), the possibility of incorporating top-down musical knowledge into transcription in a Bayesian fashion remains largely unexplored. To take one example, a frequent problem with current transcription models is that a note is often mistaken for another note an octave away (due to the similar frequency content of octave-related notes; Bello et al., 2000; Martin, 1996). Incorporating a pitch proximity factor—assigning lower probability to melodies with large leaps—should greatly alleviate this problem. There seems little doubt that such top-down knowledge could be useful in solving the engineering problem of note identification; whether it is actually used in human note-identification is an interesting question that has not yet been addressed.

6. Further issues

In this study I have proposed a model of melody perception. The model essentially incorporates just three kinds of knowledge: (a) given a key, some pitch classes are more likely than

others; (b) melodies tend to remain within a fairly narrow range; and (c) intervals between adjacent pitches tend to be small. Given this knowledge, the model is able to perform well at key detection, melodic expectation, and error detection. With regard to key finding, the model's performance is substantially better than the Longuet-Higgins/Steedman model and the Krumhansl-Schmuckler model on European folk songs; on Bach fugue themes, it outperforms the Krumhansl-Schmuckler model but not the Longuet-Higgins/Steedman model. With regard to expectation, the model performs better than both Cuddy and Lunney's (1995) model and Schellenberg's (1997) model on Cuddy and Lunney's expectation data. On balance, then—where comparison is possible—the model it at least competitive with other models in its level of performance. Beyond the issue of performance, however, I will argue here that the current model has several important advantages over others that have been proposed.

One of the important features of the model is that it is able to perform both the structural task of key identification and the surface-level tasks of expectation and error detection within a single framework. This sets it apart from prior models of both key finding and expectation, which have addressed these problems separately. It is true that the connection between key finding and expectation is indirectly reflected in some other work—notably in the fact that Krumhansl and Kessler's (1982) key profiles have been incorporated into both expectation models (Cuddy & Lunney, 1995; Schellenberg, 1997) and key-finding models (Krumhansl, 1990). But the connection between key finding and expectation is brought out much more clearly in the current model. Key finding is a matter of finding the key with which the note pattern achieves the highest joint probability; expectation is a matter of judging the probability of the note pattern itself (or the relative probability of different possible note patterns), which is the joint probability of the note pattern with all possible keys. This points to another advantage of the model: Because it offers a way of calculating the overall probability of a note pattern, it provides a method for performing error detection (and could also, potentially, contribute to the note-identification task). By contrast, regression-based models of expectation do not appear to offer any natural way of modeling error detection or pitch identification, because they provide no measure of the overall probability of a melody.

Undoubtedly, the model could be improved in many ways. This is clear when we examine its generative outputs—consider, for example, Fig. 2b, a melody generated by the model. While this melody seems much more acceptable than the one in Fig. 2a, it still does not seem very much like a real melody—even with regard to the pitch domain (the rhythmic domain is of course completely neglected). Other kinds of musical knowledge would have to be incorporated into the model to attain this goal. Melodies normally have some kind of implied harmonic structure—the notes outline a series of harmonies, forming a coherent progression and ending in a conventional closing formula or cadence; they generally also reflect some use of repeated patterns or motives. The model knows nothing of these aspects of music. It is unclear to what extent these considerations are involved in perceptual processes such as expectation and error detection, but there is some evidence that they are; Schmuckler (1989) found that listeners' musical expectancies were, indeed, affected by harmonic principles. There may also be more general "shape" considerations that could improve the model. One example is the tendency for melodies to follow small intervals with another small interval in the same direction—a phenomenon that has sometimes been called *process* (Narmour, 1990) or *inertia* (Larson, 2004). This factor is clearly evident in Cuddy and Lunney's data: for example,

given a context of an ascending major second Bb4–C4 (see Fig. 9), the highest ratings are for continuation intervals of +1, +2, or +3, creating another small interval in the same direction. This tendency is not predicted by the current model or by Schellenberg's two-factor model; it *is* predicted by the registral direction factor of Cuddy and Lunney's model, but this model performed less well than the current model or Schellenberg's model overall. Clearly, there is further work to be done in combining the strengths of these various models.

An obvious further issue that arises is the extension of the model to polyphonic music—music in which more than one note at a time is present. This raises several significant challenges. First, I have argued elsewhere (Temperley, 2004) that the method of key finding used here may not be appropriate for polyphonic music. In polyphonic music, treating each note as an independent event generated from a key profile gives too much weight to pitch classes that are repeated or doubled in different octaves; a better approach is to treat each pitch class as absent or present within a short segment of music.[16] Surface-level tasks such as error detection or expectation would also be more complex in polyphonic music. In general, polyphonic music is constructed of several simultaneous melodic lines or streams; thus, tasks such as detecting errors or generating expectations would first require that the notes be grouped into lines in the correct way. This problem of voice separation or contrapuntal analysis is a challenging problem in itself (Bregman, 1990; Temperley, 2001)—one that, incidentally, might well be susceptible to a Bayesian approach. In short, probabilistic modeling of music cognition becomes much more difficult in the case of polyphonic music, for a variety of reasons. Nevertheless, many of the problems that arise—such as polyphonic key finding and voice separation—seem amenable to probabilistic solutions.

As noted earlier, key is only one aspect of the musical structures that are inferred by listeners. Other domains of music perception have also proven to be amenable to a probabilistic approach, such as meter (Cemgil et al., 2003, 2000), phrase structure (Bod, 2002), and harmony (Raphael & Stoddard, 2004). (Probabilistic work on transcription was discussed in the previous section.) These aspects of music perception interact in complex ways with the issues of key and melodic pitch structure. One example concerns the perception of phrases. It seems clear that key plays a role in phrase perception, in that pitches that are more stable or compatible with the key are especially likely to occur at the ends of phrases (Temperley, 2001). On the other hand, rhythmic factors are also important cues to phrase structure—for example, we tend to infer phrase boundaries after long notes (Lerdahl & Jackendoff, 1983); thus, the current model on its own could not hope to detect phrase boundaries with much accuracy. An effective phrase perception model would need to take account of both pitch and rhythmic factors.[17] Another example concerns melodic expectation. While we have focused on the pitch aspect of melodic expectation, it has a rhythmic aspect as well: We form expectations not just about what note will occur but when it will occur (Jones, Moynihan, MacKenzie, & Puente, 2002). Probabilistic models of meter perception, such as that of Cemgil et al. (2003, 2000), can be used to calculate the probability of rhythmic patterns and could thus be used to model rhythmic expectation. Combining pitch and rhythm to create a complete probabilistic model of melodic expectation would seem to be an interesting possibility for the future.

In assigning a probability to any melodic input, the model described here is analogous to a language model as used in speech recognition, which assigns a probability to a sequence of words (Jurafsky & Martin, 2000). A language model can also be construed as a characterization

of actual language use—a probabilistic description of a language. Similarly, the current model might be described as a *musical style model* that characterizes a certain musical idiom with regard to principles of melodic construction. Roughly speaking, this idiom is Western tonal music, encompassing European folk music, pre-20th-century art music (classical music), and arguably much popular music as well.[18] A musical style model is successful to the extent that it assigns high probability to music within the style. (This could also be phrased in terms of the idea of cross-entropy. By having the current model assign probabilities to melodies in the Essen corpus, we are essentially measuring the cross-entropy between the model and the corpus; a better model yields lower cross-entropy.) To the extent that principles of key, range, and pitch proximity allow us to assign higher probability to a corpus of music, this provides a kind of empirical validation of these principles as claims about the music itself—and possibly about the cognitive compositional processes involved in its creation. This also raises the possibility that probabilistic musical models could be used comparatively, to characterize the differences and commonalities between musical styles. In this way, the probabilistic approach might allow for the formulation and testing of empirical claims about music, with more rigor and quantitative precision than has been possible before.

Notes

1. While the study primarily uses European folk melodies for training and testing, the basic musical principles involved are also shared with many other kinds of music—classical music, hymns, Christmas carols, nursery songs, and so on—with which most Western listeners are familiar. I return to the issue of style at the end of the article.
2. In earlier work (Temperley, 2002, 2004), I proposed a model of key detection in polyphonic music; Raphael and Stoddard's (2004) model also performs key detection as part of a harmonic analysis model. However, these models adopt a different approach to what is undertaken here; I return to this issue in section 6.
3. The Essen corpus is available at http://kern.ccarh.org/cgi-bin/ksbrowse?l=/essen/. The "europa" portion of this corpus is what was used in the current study.
4. The two melodies in Fig. 2 may be heard at http://www.theory.esm.rochester.edu/temperley/fig2a.mid and http://www.theory.esm.rochester.edu/temperley/fig2b.mid.
5. von Hippel (2000) showed that the prevalence of small intervals in melodies was not just an artifact of range constraints, by showing that scrambled versions of melodies—in which the notes were randomly reordered—had larger intervals than the original versions.
6. The model could be regarded as a Gauss-Markov model with regard to the dependency of each pitch on the previous pitch, though the dependency on the central pitch and key makes it somewhat different.
7. The key profile values could also be considered as parameters. However, to optimize these using a MLE approach—as I propose below—would be very problematic. The model has no way of figuring out the arbitrary association between keys and key profiles. Still, the model might be able to correctly sort the melodies into key categories, even if it did not associate them with keys in the correct way. This would be an interesting

experiment but will not be undertaken here. Another reason for not including the key-profile parameters in the optimization process is that it allows for fairer comparison with melodic expectation models, as discussed in section 5.

An earlier version of this model was presented in Temperley (2007). That version did not use systematic optimization techniques for setting the parameters; thus the parameter values and the test results (on the expectation and error-detection tests described below) are slightly different.

8. This optimization method appears not to have a name; it is described in Press, Teukolsky, Vetterling, and Flannery (1992, p. 413) as a simplified version of Powell's method.
9. Recall that the RPK profile values are calculated as the product of proximity profile, range profile, and key profile values, and only the key profile values depend on the key. Thus, one might wonder if key probabilities could be determined from the key profile values alone (everything else being constant across keys): that is, $P(k_x \mid \text{pitch sequence}) \propto P(k_x)\ \Pi K_n$, where K_n are the key profile values for all pitches in the melody. The problem is that the RPK values are not simply the product of the three profiles but are normalized to sum to 1; this was found to have a small effect on the model's key-finding behavior in some cases.
10. The sample was created by Paul von Hippel and was stratified to include songs from all ethnic categories represented in the Essen corpus.
11. The models of Vos and Van Geenen (1996) and Temperley (2001) are capable of identifying changes of key; for these models, the figures indicate the number of cases in which the models identified the correct initial key of the subject. The results for Longuet-Higgins and Steedman (1971) and for Vos and Van Geenen (1996) are as reported in their publications. Regarding the Krumhansl-Schmuckler model, the test reported here was done by me (using my own implementation) and simply involved giving each entire fugue subject to the model. Krumhansl (1990) also tested the Krumhansl-Schmuckler model on the Bach fugue subjects in a different way: in her test, the model was run on successively longer portions of the fugue subjects (first one note, then two notes, then three notes, etc.) until it got the right answer and then stopped. My testing method is of more relevance here, since it is the way other key-finding models have been tested and thus allows comparison.
12. Two groups of subjects were used in the experiment, a musically trained group and an untrained group. The test reported here averages the ratings for the two groups, as was apparently done by both Cuddy and Lunney (1995) and Schellenberg (1997). In the experimental stimuli, the second tone was always either C4 or F#4, but Cuddy and Lunney do not make clear which trials had C4 and which ones had F#4; in the current test, the second note was set to C4 in all cases.
13. I did not optimize the key profile values but simply used the values drawn from the Essen corpus (shown in Fig. 4). This allows fairer comparison with Cuddy and Lunney (1995) and Schellenberg (1997), since they did not include key profile values as factors in their multiple regressions.
14. Huron (2006, pp. 158–163) provides empirical data regarding scale-degree tendencies in the Essen corpus. However, his analysis focuses on the joint probability of two scale degrees occurring successively, rather than on the conditional probability of one scale

degree given another. Huron's analysis also distinguishes between different spellings of the same pitch class (e.g., #1 versus b2), which the current analysis does not.

15. In choosing a pitch for the distorted note, the correct pitch was excluded, thus ensuring that the distorted note was always different from the original note.
16. Elsewhere (Temperley, 2002, 2004) I have proposed a model of polyphonic key finding based on this idea; however, this model has no way of assessing the probability of a surface note pattern. Raphael and Stoddard's (2004) model of harmonic analysis is similar: It generates a segment of a piece simply as a "bag" of pitch class tokens, without specifying the exact register or time order of notes.
17. Bod's (2002) phrase structure model does in fact take rhythm and pitch into account. But this is done in a very data-driven way: Using a Monte Carlo approach, the model counts up all kinds of patterns of pitch and rhythm that are associated with phrase structure, such as the number of phrases that end with a half note of scale degree 1. It is possible that phrase perception could be modeled quite effectively using just a few carefully chosen pitch and rhythm parameters, similar to the approach that is taken here.
18. While we have been concerned with European folk music here, I believe that the model would, to a large extent, be applicable to other Western musical styles as well. Principles of range and proximity seem to be operative in a wide range of different styles (von Hippel & Huron, 2000). With regard to the key profiles, some modifications might be required, depending on the style. As discussed earlier, classical music seems to reflect the harmonic minor scale as the primary scale of minor keys—in which 7 is part of the scale and b7 is not—rather than the natural minor (see the profiles in Temperley, 2004, gathered from a corpus of classical music). Some popular genres also use different scalar collections, such as pentatonic or blues-based scales.

Acknowledgments

Thanks are due to Taylan Cemgil for help with the mathematical aspects of this article.

References

Aldwell, E., & Schachter, C. (2003). *Harmony and voice leading*. Belmont, CA: Thomson.

Bello, J. P., Monti, G., & Sandler, M. (2000). Techniques for automatic music transcription. In *Proceedings of the International Symposium on Music Information Retrieval*. Plymouth, MA. Retrieved from www.elec.qmul.ac.uk/people/juan/publications.htm

Bod, R. (2001). Memory-based models of melodic analysis: Challenging the Gestalt principles. *Journal of New Music Research, 31*, 27–37.

Bregman, A. S. (1990). *Auditory scene analysis*. Cambridge, MA: MIT Press.

Brown, H., Butler, D., & Jones, M. R. (1994). Musical and temporal influences on key discovery. *Music Perception, 11*, 371–407.

Carlsen, J. C. (1981). Some factors which influence melodic expectancy. *Psychomusicology, 1*, 12–29.

Cemgil, A. T., & Kappen, H. J. (2003). Monte Carlo methods for tempo tracking and rhythm quantization. *Journal of Artificial Intelligence Research, 18*, 45–81.

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.

Cohen, J. E. (1962). Information theory and music. *Behavioral Science, 7*, 137–163.

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

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

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

Deutsch, D. (1999). The processing of pitch combinations. In D. Deutsch (Ed.), *The Psychology of Music* (pp. 349–411). San Diego, CA: Academic Press.

Eisner, J. (2002). Discovering syntactic deep structure via Bayesian statistics. *Cognitive Science, 26*, 255–268.

Godsmark, D., & Brown, G. J. (1999). A blackboard architecture for computational auditory scene analysis. *Speech Communication, 27*, 351–366.

Hiller, L. A., & Fuller, R. (1967). Structure and information in Webern's Symphonie, Opus 21. *Journal of Music Theory, 11*, 60–115.

Huron, D. (1999). *Music research using Humdrum: A user's guide*. Retrieved from http://dactyl.som.ohio-state.edu/Humdrum/guide.toc.html

Huron, D. (2006). *Sweet anticipation*. Cambridge, MA: MIT Press.

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

Juliano, C., & Tanenhaus, M. K. (1994). A constraint based lexicalist account of the subject/object attachment preference. *Journal of Psycholinguistic Research, 23*, 459–471.

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing*. Upper Saddle River, NJ: Prentice-Hall.

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

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge: Cambridge University Press.

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

Krumhansl, C. L., Louhivuori, J., Toiviainen, P., Järvinen, T., & Eerola, T. (1999). Melodic expectation in Finnish spritiual folk hymns: Convergence of statistical, behavioral, and computational approaches. *Music Perception, 17*, 151–195.

Lake, W. (1987). *Melodic perception and cognition: The influence of tonality*. Unpublished doctoral dissertation, University of Michigan.

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

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

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

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

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*, 676–703.

Manning, C. D., & Schütze, H. (2000). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Martin, K. (1996). *A blackboard system for automatic transcription of simple polyphonic music* (M.I.T. Media Laboratory Perceptual Computing Section Tech. Rep. No. 385).

Meyer, L. B. (1956). *Emotion and meaning in music*. Chicago: University of Chicago Press.

Meyer, L. B. (1973). *Explaining music*. Berkeley: University of California Press.

Miller, G. A., & Heise, G. A. (1950). The trill threshold. *Journal of the Acoustical Society of America, 22*, 637–638.

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

Olman, C., & Kersten, D. (2004). Classification objects, ideal observers and generative models. *Cognitive Science, 28*, 227–239.

Osherson, D. (1990). Judgment. In D. Osherson & E. Smith (Eds.), *An invitation to cognitive science, Vol. 3: Thinking* (pp. 55–87). Cambridge, MA: MIT Press.

Petersen, K. B., & Petersen, M. S. (2005). *The matrix cookbook*. Lyngby, Denmark: Technical University of Denmark. Retrieved from http://www2.imm.dtu.dk/pubdb/p.php?3274

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

Povel, D.-J. (1996). Exploring the fundamental harmonic forces in the tonal system. *Psychological Research, 58*, 274–283.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in C: The art of scientific computing*. Cambridge, UK: Cambridge University Press.

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

Raphael, C. (2002b). Automatic transcription of piano music. *Proceedings of the 3rd Annual International Symposium on Music Information Retrieval*. Retrieved from http://xavier.informatics.indiana.edu/~craphael/papers/index.html

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

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*, 27–52.

Schaffrath, H. (1995). *The Essen Folksong Collection in the Humdrum kern format*. D. Huron (Ed.). Menlo Park, CA: Center for Computer Assisted Research in the Humanities.

Schellenberg, E. G. (1996). Expectancy in melody: Tests of the implication-realization model. *Cognition, 58*, 75–125.

Schellenberg, E. G. (1997). Simplifying the Implication-Realization model of melodic expectancy. *Music Perception, 14*, 295–318.

Schellenberg, E. G., Adachi, M., Purdy, K. T., & McKinnon, M. C. (2002). Expectancy in melody: Tests of children and adults. *Journal of Experimental Psychology: General, 131*, 511–537.

Schmuckler, M. (1989). Expectation and music: Investigation of melodic and harmonic processes. *Music Perception, 7*, 109–150.

Sloboda, J. A. (1976). The effect of item position on the likelihood of identification by inference in prose reading and music reading. *Canadian Journal of Experimental Psychology, 30*, 228–236.

Sobel, D. M., Tenebaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28*, 303–333.

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

Temperley, D. (2002). A Bayesian approach to key-finding. In C. Anagnostopoulou, M. Ferrand, & A. Smaill (Eds.), *Music and artificial intelligence* (pp. 195–206). Berlin: Springer.

Temperley, D. (2004). Bayesian models of musical structure and cognition. *Musicae Scientiae, 8*, 175–205.

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

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems* (vol. 11, pp. 59–65). Cambridge, MA: MIT Press.

Thompson, W. F., Cuddy, L. L., & Plaus, C. (1997). Expectancies generated by melodic intervals: Evaluation of principles of melodic implication in a melody-completion task. *Perception & Psychophysics, 59*, 1069–1076.

Unyk, A. M., & Carlsen, J. C. (1987). The influence of expectancy on melodic perception. *Psychomusicology, 7*, 3–23.

von Hippel, P. (2000). Redefining pitch proximity: Tessitura and mobility as constraints on melodic intervals. *Music Perception, 17*, 315–327.

von Hippel, P., & Huron, D. (2000). Why do skips precede reversals? The effect of tessitura on melodic structure. *Music Perception, 18*, 59–85.

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

Youngblood, J. E. (1958). Style as information. *Journal of Music Theory, 2*, 24–35.