Probabilistic Models of Melodic Interval

David Temperley
Eastman School of Music, University of Rochester

Two probabilistic models of melodic interval are compared. In the *Markov model*, the "interval probability" of a note is defined by the corpus frequency of its melodic interval (the interval to the previous note), conditioned on the previous one or two intervals; in the *Gaussian model*, the interval probability is a simple mathematical function of the size of the note's melodic interval and its position in relation to the range of the melody. In both models, this interval probability is then multiplied by the probability of the note's scale degree to yield its actual probability. The two models were tested on four corpora of tonal melodies using cross-entropy. The Markov model yielded a somewhat lower (better) cross-entropy than the Gaussian model, but is also much more complex, requiring far more parameters. The models were also tested on melodic expectation data, and on their ability to predict the distribution of intervals in a corpus. Possible ways of improving the models are discussed, as well as their broader implications for music cognition.

Received: October 29, 2013, accepted February 8, 2014.

Key words: melody, cross-entropy, Markov models, corpus research, expectation

Probabilistic modeling has lately become an important and influential approach in the field of music cognition, as it has throughout cognitive science. Probabilistic methods have been successfully applied to a number of problems in music perception, such as rhythm perception (Sadakata, Desain, & Honing, 2006), key-finding (Temperley, 2007), harmonic analysis (Raphael & Stoddard, 2004), and segmentation (Bod, 2002; Pearce & Wiggins, 2012). The probabilistic approach has a number of virtues: it can successfully handle ambiguous input, and can reflect the ambiguous or uncertain mental representation of such input; it can represent perceptual constraints or preferences of a gradient nature; and it can simulate the process whereby perception is shaped by regularities in the environment. Probabilistic methods have proven useful not only for modeling cognition but also for practical problems of music information retrieval, such as style identification (Chai & Vercoe, 2001) and transcription (Klapuri & Davy, 2006).

While the problems mentioned in the previous paragraph relate to the perception and processing of music, probabilistic methods have also been applied to the modeling of music itself (Conklin & Witten, 1995; Mavromatis, 2005; Pearce & Wiggins, 2004; Temperley, 2010). A probabilistic model can be used to assign a probability to a piece of music, or some aspect of it, such as a sequence of pitches or durations. This is useful in at least two ways. First, it provides a way of modeling the cognitive processes of composition—the factors and procedures involved in the creation of music. Under certain assumptions (as discussed further below), the probability assigned by a model to a body of data can be taken as an indicator of the probability of the model given the data. Competing models of the compositional process can therefore be evaluated by the probabilities that they assign to the music under investigation. No probabilistic model (to my knowledge) has claimed to incorporate *all* the factors involved in the compositional process; but given one or more factors that are known to play a role in that process, probabilistic methods can be used to determine the best (most cognitively plausible) way of modeling them. A second use of probabilistic models of music is in the modeling of expectation. It seems reasonable to suppose that, when listeners form expectations for the continuation of a musical context (such as a melody), they are evaluating the probabilities of possible continuations, in combination with the previous context. Probabilistic models can be used to simulate this process.

Most probabilistic models of music have focused on melody—the assignment of probabilities to a single sequence of notes—and I will also do so here. While probabilistic models of melody have considered a variety of musical dimensions, two factors have emerged as particularly important. One is *interval*: the probability of note within a melodic line depends in part on the resulting interval in relation to the previous note. The other is *scale degree*, that is, pitch-class in relation to the tonic: some scale degrees are more probable than others, based on their position in the scale of the relevant tonal system. Most models of melody—probabilistic or

Music Perception, volume 32, issue 1, pp. 85–99, issn 0730-7829, electronic issn 1533-8312. DOI: 10.1525/mp.2014.32.1.85

not—have incorporated these two factors in some form. (One well-known exception is the Implication-Realization theory of Narmour [1990], which focuses on melodic shape rather than tonal factors. But even Narmour acknowledges the importance of tonal factors, and those implementing his theory, such as Schellenberg [1996], have found it necessary to include them.)

The current study compares two models reflecting alternative approaches to the modeling of melodic interval; I will call these the *Markov* model and the *Gaussian* model. In the Markov model, the probability of an interval is determined by its observed frequency in a corpus, perhaps conditional on the previous one or more intervals. The Markov model represents a modeling paradigm known as the Markov approach or *n-gram* approach. This approach has been highly influential in recent music research, most notably in the context of multiple-viewpoint models, a strategy pioneered by Conklin and Witten (1995) and further developed by Pearce and Wiggins (2004, 2006). In a multiple-viewpoint system, the probabilities for individual features of a note (perhaps conditioned on the features of previous notes) are combined to yield a single probability for the note. Multiple-viewpoint studies have often included interval and scale degree as features, along with numerous others such as absolute pitch, contour, rhythmic attributes, and more complex features such as interval to the first note of the measure.

The second, *Gaussian,* model, is also probabilistic. In this model, however, the probabilities of intervals are defined not by their frequency in a corpus, but rather, by the general principle of *pitch proximity*: each pitch in a melody tends to be close to the previous pitch, so that small intervals are more frequent than large ones. This is a well-known principle of music theory (often reflected in compositional teaching, e.g., Gauldin 1985, p. 17; Aldwell & Schachter, 2003, p. 69) and auditory psychology (Deutsch, 1999); it has also been confirmed statistically in a variety of musical styles (Huron, 2006; von Hippel, 2000). Thus the probability of an interval can be defined as a simple mathematical function of its size; in the current model, a Gaussian function (i.e., a normal distribution) is used. The probability of an interval appears to depend on its context in some ways; for example, a large interval is usually followed by an interval in the opposite direction—a phenomenon known as "post-skip reversal." Narmour's Implication-Realization theory (1990) includes an explicit preference for large intervals to be followed by a change of direction, and quantifications of the theory have included this preference as well (Cuddy & Lunney, 1995; Schellenberg, 1996, 1997). It has also been suggested, however, that post-skip reversal may arise simply from constraints on range; a large interval will tend to take a melody near the edge of its range, thus naturally favoring a return to the center (von Hippel & Huron, 2000). The Gaussian model presented here reflects this latter approach, assigning higher probabilities to notes that stay within the previously established range.

Most previous models of melody have found that it is desirable to incorporate scale-degree information in some way, and I will do so here. Scale-degree probabilities are simply modeled in a statistical fashion based on their corpus frequency, in both the Markov and Gaussian models. The possibility of a Markov approach to scale degree (conditioning scale degrees on previous scale degrees) will also be considered.

Figure 1 shows the distribution of melodic intervals in the Essen Folksong Collection, a corpus of over 6,000 European (mostly German) melodies (Schaffrath, 1995). Log (base 2) probabilities are used, to bring out distinctions between small values. This figure gives insight into the motivation for both the Markov and Gaussian approaches to melodic interval. The principle of pitch proximity is clearly reflected, in that larger intervals are generally less frequent than smaller ones. There are also many local ups and downs in the distribution; for example, whole steps (+2 and -2) are more than twice as common as half steps (+1 and -1), despite being larger. These local fluctuations might seem difficult to capture by any simple rule, thus favoring a Markovian approach that represents the frequency of each interval. It is also important, however, to consider the role of scale degree. It is a well-known fact that chromatic intervals vary in their frequency within the diatonic scale: considering just a one-octave range (C to C) in C major, there are five whole steps (C-D, D-E, F-G, G-A, A-B) but only two half steps (E-F, B-C). Thus, given that diatonic scale degrees tend to be more frequent than non-diatonic ones, the greater frequency of major over minor seconds may be due, at least in part, to the fact that stepwise motion along the diatonic scale is more likely to give rise to major seconds than minor seconds. Similar reasoning might be applied to other irregularities in the interval distribution. One of the goals of the current study is to investigate how well the distribution of intervals in tonal music can be explained by the principle of pitch proximity in combination with a preference for certain scale degrees over others.

In what follows, the Markov and Gaussian models are tested on their ability to predict three kinds of data: (1) sequential data from melodic corpora, (2) experimental data from melodic expectation studies, and (3) the distribution of melodic intervals shown in Figure 1. I begin

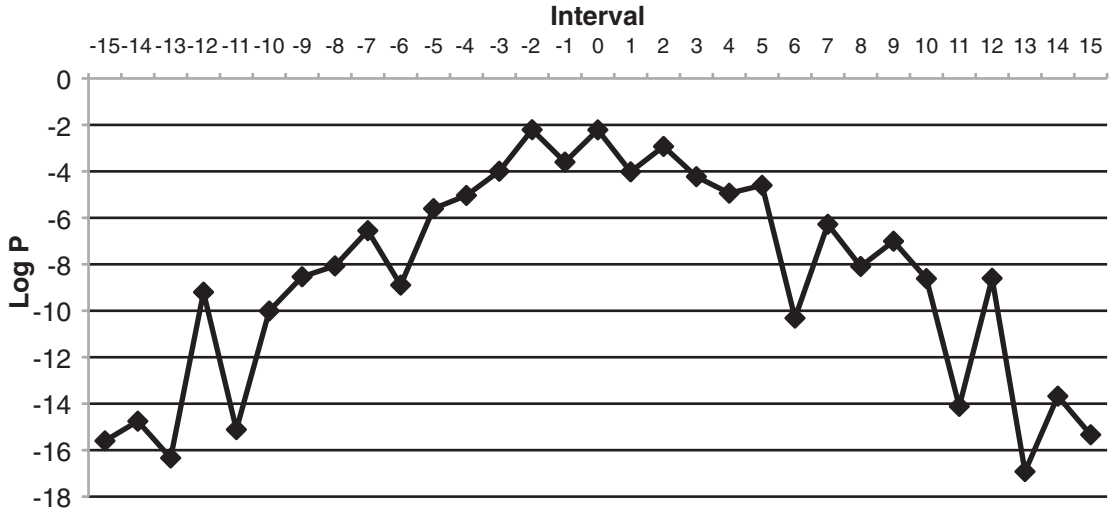

FIGURE 1. The distribution of melodic intervals in the Essen Folksong Collection. Probabilities are shown in log base 2.

by describing the models in more detail and explaining the general testing strategy.

THE MODELS AND THE TESTING STRATEGY

The input to both models is simply a sequence of pitches, in integer notation (middle C = 60, etc.). No rhythmic information is included. The models also require scale degree information; that is, the relationship of each note to the tonic. A simple way to provide this is to transpose all melodies down to the key of C, so that, for example, pitch 60 (plus or minus any multiple of 12) is always scale degree 1. (Major scale degrees will be represented here by integers 1 through 7, and other degrees as altered versions of these, using the most common spelling of each: #1, b3, #4, b6, and b7.) The corpora used for training and testing all indicate the key of each piece, so this can easily be done. (For modeling expectation data, a different approach is required, as will be explained later on.) Since the models considered below make no use of absolute pitch information, it makes no difference whether the melodies are transposed up or down. The input also indicates whether each melody is in a major or minor key (MA or MI is inserted at the beginning of each melody); how this information is used will be explained below.

The Markov model computes a probability for each note based on its scale degree and its interval to the previous note. It begins by computing separate probabilities for the interval and the scale degree. Each of these two probabilities is computed in a Markovian fashion, conditioned on features of the previous notes—intervals or scale degrees, respectively. The order of the Markov model—that is, the number of previous events on which each event is conditioned—may be zero, one, or two; all of these possibilities will be considered. These are known as *zeroth-order*, *first-order*, and *second-order* models, respectively. They are also known as *unigram*, *bigram*, and *trigram* models; these terms refer to melodic sequences of one, two, or three elements, and thus indicate the size of the melodic units ("n-grams") being counted. Higher-order n-gram models are also possible, but will not be considered here.

The model's probabilities are set by the counts of events and contexts in a corpus. For example, if an interval of -2 occurs 100 times in a corpus and is followed by an interval of -1 on 20 of those occurrences, the first-order probability of -1 given a previous interval of -2 would be 20 / 100 = .2. Given the interval and scale-degree probabilities, the model then simply multiplies them together. In effect, this favors pitches whose interval and scale-degree probabilities are both relatively high. The resulting values are not true probabilities, as the values for all possible pitches in a given context do not sum to 1; the values are divided by a normalizing constant (the sum of all of them) so that they do sum to 1, and the value for the pitch in question is the probability assigned to it. This model essentially adopts the approach of multiple-viewpoint modeling (Conklin & Witten, 1995), and may be regarded as a simple model of that kind; it is not, to my knowledge, identical to any model that has been specifically proposed previously, though some studies have considered many combinations of features and may well have considered this one. Combining distributions by multiplying them

together is a standard approach in multiple-viewpoint modeling (e.g., Pearce, Conklin, & Wiggins, 2005; Pearce & Wiggins, 2006), though other approaches have also been used.

The Gaussian model, likewise, computes a probability for each note based on its scale degree and intervallic context, but does so in a rather different way. The probability of an interval is computed using two functions. One is a *proximity profile*, a normal distribution centered around the previous pitch; another is a *range profile*, also a normal distribution, centered around the mean pitch of all the preceding notes in the melody (this is taken as an approximation of the range of the melody). The interval probability of a pitch is given by its position in these two normal distributions, multiplied together. Multiplying two normal distributions creates a third normal distribution whose mean is between the means of the distributions being multiplied; the effect of this is to favor notes that are close to both the previous pitch and the center of the melody's range. This interval probability value is then multiplied by the scale-degree probability of the note, which is simply the zeroth-order Markov probability of that scale degree. (I will refer to a zeroth-order scale-degree distribution a "scale-degree profile.") Formally:

$$P(E_n) = N(E_n; E_{n-1}, v_p) \times N(E_n; M_n, v_r) \times SD(E_n)/Z \quad (1)$$

where E_n is a possible pitch for the nth note of the melody; $N(E_n; m, v)$ refers to the value of E_n in a normal distribution with mean m and variance v; $E_{n\text{-}1}$ is the previous pitch; M_n is the mean pitch of the melody up to (but not including) the current note; v_p and v_r are the variances of the proximity and range profiles; $SD(E_n)$ is the value of the current pitch in the scale-degree profile; and Z is a normalizing constant, to ensure that the values for all possible pitches sum to 1. In effect, this favors pitches that are close to the previous pitch, close to the mean pitch of the melody, and probable in terms of their scale degree. This model is essentially that proposed in Temperley (2007, 49–64; 2008); the main difference is that the mean of the range profile in that earlier work was computed in a more complex way (taking into account that it is likely to be near a certain point in absolute terms).

Given these models, with the necessary parameters defined, it is straightforward to compute the probability of a note in a given context. The probability of a melody, or indeed an entire corpus, can then be computed as the product of the probabilities of all the notes. Since the resulting probabilities can be very small numbers, it is convenient to represent them by logarithms. Dividing the total value for the corpus by the number of notes produces the "per-note log probability" of the data given the model; since the log of a probability is always negative, we add a negative sign to make it positive. This produces a quantity known as *cross-entropy*:

$$\text{cross-entropy} = (-1/N) \sum_n \log_2 P_m(E_n) \quad (2)$$

where N is the number of events in the corpus and $P_m(E_n)$ is the probability assigned by the model to the nth event.

Cross-entropy—representing the probability that a model assigns to a body of data—gives an indication of how well the model fits or predicts the data, and thus, how good it is as a model of the underlying process that gave rise to the data. Several prior studies have used cross-entropy to evaluate models of melody (Conklin & Witten, 1995; Pearce & Wiggins, 2004; Temperley, 2010). The logic behind this approach is simple and elegant. If two models are equal in prior probability—that is, if they seem equally probable before the data is seen—it follows mathematically that the one yielding lower cross-entropy (higher probability) is the more probable one given the data. From Bayes' rule:

$$P(\text{model} \mid \text{data}) = P(\text{data} \mid \text{model})P(\text{model})/P(\text{data}) \quad (3)$$

P(data), the overall probability of the data in combination with all possible models, is the same for any given model, so

$$P(\text{model} \mid \text{data}) \propto P(\text{data} \mid \text{model})\text{P}(\text{model}) \quad (4)$$

where "$\propto$" means "is proportional to." And if the prior probability P(model) is the same for all models:

$$P(\text{model} \mid \text{data}) \propto P(\text{data} \mid \text{model}) \quad (5)$$

Of course, models may not be equal in prior probability; in the current case, there may well be other considerations (such as historical or psychological evidence about the compositional process) that lead us to favor one model over another *a priori*. Another general consideration that may affect the prior probability of a model is simplicity; other things being equal, a simpler model is generally considered more plausible. The number of parameters required by a model is often taken as a measure of its complexity (Akaike, 1974; Mavromatis, 2005; Schwarz, 1978), and I will do so here. In computational terms, the kind of complexity at issue here is *space complexity*, the amount of memory needed (Hartmanis & Stearns, 1965); this should not be confused with *time*

complexity, the amount of computing time required, which will not be considered here.

There is frequently a trade-off between simplicity and goodness-of-fit; a model requiring more parameters will often fit the data better (though not always). But there is no objective, widely accepted method of balancing simplicity against goodness-of-fit; how one does this may well depend on the specific purpose for which the model is being used. In what follows, I report goodness-of-fit and complexity measures for each model, and consider one possible way of combining them into a single measure; but ultimately, I leave it to the reader to decide how to balance these criteria.

A study such as this requires training data, to set the parameters of the models. These parameters include the scale-degree probabilities of both models and the interval probabilities for the Markov model. The variances for the Gaussian model's normal distributions must also be set (a higher variance creates a flatter distribution). When testing on corpus data, it is appropriate that the training data be stylistically similar to the test data, but it is important to use different training data from that used for testing. We wish to prevent the model from "overfitting"—giving high probabilities to events that happen by chance to occur in the training set; testing the model on its training data rewards overfitting. Here I adopt the method of cross-validation. Each corpus is divided into ten equal portions, with nine of the portions used for training and the remaining one for testing; the process is repeated ten times, using a different portion for testing each time, and the model's score is the average cross-entropy across the ten test sets.

The cross-entropy approach gives us a method of testing our models on corpus data. For expectation data, a somewhat different approach is required. Here the data to be modeled are human judgments of the expectedness of notes in a given context; we evaluate a model on how well the probabilities it assigns to the notes match the expectedness judgments. This will be explained in greater detail below.

Certain aspects of each model were varied, with the aim of finding the optimal performance of each model. For the Markov model, both the interval and scale-degree probabilities can be computed using zeroth-, first-, or second-order probabilities; different orders for the two components were also considered, creating nine possible combinations. The probabilities for all of these versions of the model were set from training data, by counting interval or scale-degree unigrams, bigrams, or trigrams, as appropriate. For the Gaussian model, the variances of the two normal distributions were optimized. The logical way of doing this is by choosing the values that yield lowest cross-entropy on the training data. This was tried with different training sets; using different training portions within the same corpus always yielded the same optimal values (only integer values were tried). Thus, for each corpus, the optimal values were determined from one training portion and these were used for all the tests on that corpus.

One problem is how to handle events seen in testing that have never been encountered in training. Strictly speaking, the probability of these events should be zero, but in that case their negative log probability (and therefore that of the whole corpus) would be infinite. This is mainly a problem for the Markov model, since there may well be interval trigrams (or even bigrams or unigrams) seen in testing that were not encountered in training; a similar problem may occur with scale-degree n-grams. (It is not a problem for the Gaussian model, as long as each scale degree has been encountered at least once.) Here I adopt an extremely simple solution: when a trigram seen in testing has not been seen in training, the model uses (or "backs off" to) the bigram probability instead; when a bigram has not been seen, it backs off to the unigram; when an interval unigram has not been seen (which is extremely rare), it backs off to the octave. This method is not strictly legitimate, since the probability of an event is only set once it is seen, and the probabilities of all possible pitches calculated in this way may sum to slightly more than 1. In effect, this approach is over-generous to the Markov model, slightly overstating the true probability that it assigns to the corpus. Because of the large size of the training sets, however, backoff rarely occurred: with the Essen corpus, for example, only 1.2% of the interval trigram tokens encountered in testing were unseen in training. Alternative backoff methods were also tried (such as computing each trigram probability as a weighted sum of observed trigram and bigram probabilities), and it was found that this resulted in very little difference in cross-entropy.

Another problem for both models is what to do at the beginning of the melody. Even the zeroth-order Markov model requires at least one previous note to define the interval of the current note; the second-order version of the model requires three preceding notes. The Gaussian model, too, requires preceding notes for the proximity and range profiles to be defined. There are good ways of solving this problem; one could, for example, define the probability of the first note with a distribution over absolute pitches rather than intervals. It seemed unlikely that adding this feature would greatly alter the relative performance of the models. In the interest of simplicity, I evade the problem by skipping the first three notes of

each melody in the cross-entropy calculations (but including them in the contexts for following notes, where necessary).

One final issue concerns major and minor keys. The reason for incorporating scale-degree profiles into the models is that the probability of a note depends on its position in the current scale. But the current scale depends not only on the tonic but also on whether the key is major or minor; therefore a probabilistic model might perform better if it possessed this information and applied a major or minor scale-degree profile accordingly. It is not obvious that this would improve performance; dividing the data into major and minor portions leaves less training data for each scale-degree profile, which may result in less optimal values. But experiments (not reported in detail here) showed that both the Markov model and the Gaussian model yielded better performance when major and minor melodies were separated; therefore this was done in the tests reported below.

TESTING ON SEQUENTIAL CORPUS DATA

Four corpora of melodies were used to test the two models: (1) The *folksong corpus* consists of 6,208 songs from the Essen Folksong Collection (Schaffrath, 1995). (2) The *chorale corpus* consists of 159 Bach chorale melodies.[1] (3) The *classical corpus* contains 9,788 instrumental melodies from Barlow and Morgenstern's (1948) *Dictionary of Classical Themes*, encoded in Humdrum notation by David Huron. (4) The *rock corpus* consists of 162 melodies from songs on *Rolling Stone* magazine's list of the "500 Greatest Songs of All Time" (*Rolling Stone*, 2004; Temperley & de Clercq, 2013). (The entire rock corpus contains 200 melodies; songs with modulations were excluded, as were songs containing no melodic information, such as rap songs.) In the folksong, chorale, and classical corpora, melodies are labeled with major and minor keys; different scale-degree profiles for major versus minor melodies were learned in training and applied in testing. In the rock corpus, the melodies are labeled with tonal centers but not with major and minor (since this distinction is problematic in rock), thus a single set of scale-degree probabilities was applied to all songs.

Both the chorale corpus and parts of the folksong corpus have been used in other melody-modeling research. Conklin and Witten (1995) measured cross-entropy on a set of five chorale melodies, using a multiple-viewpoint system with a variety of features, achieving a per-note cross-entropy of 1.87. Pearce and Wiggins (2004, Table 8) used the entire chorale corpus and parts of the Essen Folksong Collection, using a Markov model that considered only absolute pitch; they achieved cross-entropies of 2.35 for the chorale corpus and values between 2.11 and 2.69 for different portions of the Essen corpus. Comparison of those models with the current ones is difficult, due to the many differences; as stated earlier, my aim is to compare approaches to the modeling of interval in a controlled fashion.

For the Markov model, each corpus was tested with zeroth-order, first-order, and second-order versions of the interval and scale-degree components, in all combinations. For the scale-degree component, a uniform distribution was also tried, assigning equal probabilities to all scale degrees; in effect, this version of the model calculates probabilities by interval alone. (It is not possible for the model to calculate probabilities by scale degree alone, since this specifies only pitch-class, not pitch.) Table 1 shows the complete results of this for the Essen corpus. Each test result indicates the cross-entropy, that is, the per-note negative log probability that the model assigns to the data. It can be seen that the best performance (lowest cross-entropy) occurs with a second-order interval model and a zeroth-order scale-degree model. The Gaussian model was tested with different combinations of proximity variance and range variance (trying integer values only). Table 2 shows, for all four corpora, the best version (combination of scale-degree and interval orders) of the Markov model, the best variance values for the Gaussian model, the cross-entropy for each of the two models, and the ratio between their cross-entropies. For the Markov model, the table shows that the combination of a second-order interval model and a zeroth-order scale-degree model is optimal for all four corpora; considering bigram and trigram scale-degree probabilities yielded no benefit.

On all four corpora, the cross-entropy of the Markov model is lower than that of the Gaussian model. The

TABLE 1. *Cross-Entropy Values for the Markov Model on the Folksong Corpus for Different Orders of Interval and Scale-Degree.*

	Scale-degree order			
Interval order	Uniform	Zeroth	First	Second
Zeroth	3.35	2.85	2.91	2.78
First	2.93	2.67	2.74	2.75
Second	2.69	**2.56**	2.65	2.68

[1] Both the folksong and chorale corpora were acquired from the Musedata archive at the Center for Computer Assisted Research in the Humanities (*www.musedata.org*). The complete corpus available at the archive contains 185 chorales, but this includes 26 duplicate melodies which were removed.

TABLE 2. *Tests of the Markov and Rule-based Models on Four Corpora.*

Corpus	Num. notes in test set[1]	Markov model (MM)			Gaussian model (GM)			GM/MM cross-entropy	GM/MM AIC
		Optimal orders (sc.-deg, interval)	Num. params.[2]	Cross-entropy	Optimal variances (range, proximity)	Num. params.	Cross-entropy		
Folksong	28105	0, 2	15649	2.56	23, 11	26	2.73	1.06	0.81
Classical	15323	0, 2	15649	2.97	50, 15	26	3.36	1.13	0.76
Chorale	702	0, 2	15649	2.26	19, 7	26	2.62	1.16	0.08
Rock	4821	0, 2	15637	2.78	34, 11	14	2.98	1.07	0.40

Note: [1]Each test set is 10% of the entire corpus. [2]Very conservative estimates, assuming no parameters for intervals larger than an octave.

difference varies between 6% and 16%. But simplicity must also be considered. This raises the question of how many parameters are required by each model. For an alphabet of symbols of size N, a zeroth-order model requires N parameters, a first-order model requires N^2, and a second-order model requires N^3. For scale degrees, the alphabet size is 12; there are two scale-degree profiles (major and minor), thus 24 parameters are required for both models (assuming zeroth-order scale-degree probabilities for the Markov model). For intervals, the issue is more difficult. In theory, an interval of any size might occur. This is not a problem for the Gaussian model; the probability for any interval can easily be calculated from the normal distributions. The Markov model, however, needs parameters for each interval (or combination of intervals, for first- and second-order models). The model should at least allow all intervals up to the largest interval present in the data. For the Essen collection, the largest interval is 21 (an octave plus a major sixth); this creates a total interval range of 43 (21 ascending + 21 descending + repetition), requiring $43^3 = 79507$ parameters. One could perhaps devise a simpler scheme for handling very large intervals—for example, assigning them all the same very low probability. If we assume this approach for all intervals greater than an octave (though this was not actually done in implementing the model), this still yields $25^3 = 15625$. Thus:

> Total parameters for the Gaussian model = 24 (scale degrees) + 1 (proximity profile variance) + 1 (range profile variance) = 26
>
> Total parameters for Markov model = 24 (scale degrees) + 15625 (intervals) = 15649

By this (very conservative) estimate of the number of parameters in the Markov model, it requires 602 times as many parameters as the Gaussian model. This applies to the folksong, classical, and chorale corpora; for the rock corpus, we subtract 12 from each total since only one scale-degree profile is used.

Several proposals have been made for how the criteria of complexity and goodness-of-fit might be balanced against one another. One is the Akaike Information Criterion (AIC) (Akaike, 1974). According to this metric, a model is evaluated by the following formula:

$$\mathrm{AIC} = 2k - (2 \times \ln P(\text{data} \mid \text{model})) \qquad (6)$$

where k is the number of parameters in the model; ln P(data | model) can be calculated as the cross-entropy multiplied by the number of notes in the test set (multipled by a constant to convert log base 2 to natural logarithms). A lower AIC means a better model. The ratio between the Gaussian model's AIC and the Markov model's AIC for each corpus test is shown in the rightmost column of Table 2; this measure favors the Gaussian model on all four corpora. (The weight of goodness-of-fit in relation to simplicity increases as the size of the test set increases, meaning that the advantage of the Gaussian model is greater for smaller corpora.) However, the AIC is not universally accepted; alternative ways of balancing goodness-of-fit and simplicity have also been proposed (Pitt, Myung, & Zhang, 2002; Rissanen, 1989; Schwarz, 1978).

Several features of Table 2 deserve further discussion. Regarding the Markov model, the fact that a zeroth-order scale-degree model is optimal for all four corpora is worthy of comment. Conventional wisdom holds that certain scale-degree patterns are particularly common in tonal music, such as $\hat{7}$-$\hat{1}$ and $\hat{4}$-$\hat{3}$ (Aldwell & Schachter, 2003, p. 91). One might suppose that a first-order (or higher) scale-degree model would be needed to capture such patterns. The current study suggests, however, that they may arise naturally from more general principles—an intervallic preference for half-step motion, combined with zeroth-order preferences for scale degrees $\hat{1}$ and $\hat{3}$. Indeed, the fact that the Gaussian model fits the data nearly as well as the Markov model suggests that even the apparent preference for half-step motion may simply reflect a still more general preference for small melodic intervals.

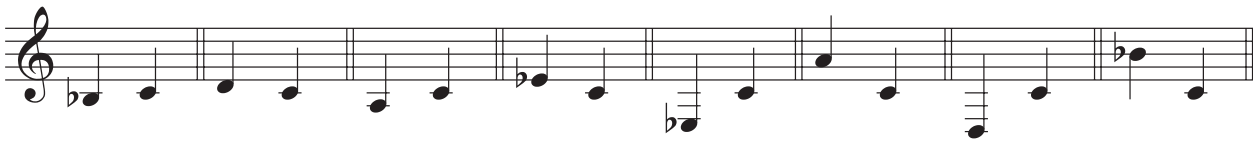

FIGURE 2. The two-note contexts used in Cuddy and Lunney's (1995) experiment.

Table 2 also reflects some interesting differences between the corpora. Regarding the Gaussian model, the optimal range and proximity variances are somewhat greater for the classical corpus than for the other three. No doubt this is partly because the classical corpus contains instrumental themes while the other corpora feature vocal melodies; instrumental melodies often feature larger intervals and ranges than vocal ones. The classical corpus also reflects higher cross-entropy (by both models) than the other three. This may be due, again, to the larger range of intervals used in the classical corpus, and more generally to its stylistic diversity. The Barlow and Morgenstern dictionary includes themes from the 17th through the 20th century, and thus spans a range of melodic idioms; for example, it contains many highly chromatic themes (even a few truly atonal ones) as well as many that are purely diatonic. The stylistic heterogeneity of the classical corpus may also explain why the difference in cross-entropy beween the two models is relatively large in this case. The second-order interval component of the Markov model gives it a (very limited) ability to adjust to stylistic context; it might recognize, for example, that a context of two large intervals suggests a rather "leapy" idiom, which is likely to be followed by another large interval. This reasoning does not, however, explain the even larger difference between the models on the chorale corpus, which one might suppose was the most stylistically homogenous of the four. A possible explanation in this case might be that the chorale corpus features certain melodic idioms that tend to occur often; such patterns could be learned by the second-order interval component of the Markov model, thus giving it an advantage over the Gaussian model. This hypothesis is supported by the fact that the Markov model assigns the chorale corpus lower cross-entropy than any of the other three. Informal inspection of the chorale corpus offered little support for this view, however; while it contains many occurrences of intervallic and scale-degree patterns typical of common-practice Western music, such as $\hat{3}$-$\hat{2}$-$\hat{1}$, such patterns appear to be almost equally common in the folksong and classical corpora. Thus, the Markov model's especially low cross-entropy on the chorale corpus remains something of a mystery.

TESTING THE MODELS ON EXPECTATION DATA

A model that assigns a probability to a sequence of notes can also be used to model human expectation judgments. A large body of work has been devoted to the modeling of musical expectations, mostly focusing on the dimension of pitch (for a survey, see Temperley, 2012). One particularly widely used data set was created by Cuddy and Lunney (1995). In this study, participants heard two-note contexts followed by a continuation note that could be anywhere within an octave of the second context note (thus 25 different continuation notes were possible for each context); eight different contexts were used, as shown in Figure 2, creating 200 stimuli in all. Participants rated the expectedness of the continuation note given the two-note context on a scale of 1 to 7. Several attempts have been made to model this data. Cuddy and Lunney (1995) and Schellenberg (1996, 1997) used models inspired by Narmour's (1990) theory of melody, in which the predicted expectedness of a continuation was a function of several principles of melodic shape; scale degree was also included as a factor. Multiple regression was used to find the optimal fit of the factors to the data, yielding a correlation of $r = .80$ in Cuddy and Lunney's study and $r = .85$ in Schellenberg's. Pearce and Wiggins (2006) modeled the data using a multiple viewpoint approach. They considered various factors relating to interval and melodic shape, but not scale degree; they also considered rhythm, unlike the other models discussed here. (Rhythm is informative since the second context note was shorter than the other two.) The best version of the model combined three features: interval to the first note of the pattern, diatonic interval, and a combination of interval and duration. The model's probabilities were transformed into rankings and yielded a correlation of .85. Finally, Temperley (2008) applied a probabilistic model very similar to that proposed above, though with some additional features added (as will be explained below), achieving a correlation of .88.

My aim here is not to improve on these models in terms of their fit to the data, but rather, to address the same question asked in the previous section: Given a model that considers only interval and scale degree, is a Markovian approach to interval better than a Gaussian

one? In what follows, I use the two models presented earlier to predict Cuddy and Lunney's (1995) data. Each model yields a log probability for each note in a melody given the previous context; these values can be treated as the models' expectation judgments and compared with the participants' judgments in Cuddy and Lunney's experiment.

One problem arising here concerns key. In the earlier tests on corpora, the key of each melody was known and was used to transpose all melodies to the same key, so the scale degree of each note could be used in testing. With expectation data, however, the key may be unknown or ambiguous, especially with short melodic segments such as Cuddy and Lunney's (1995) three-note patterns. The modeling approach used here allows an effective solution to this problem. Each note probability emitted by a model indicates $P(E_n \mid \text{context, key})$, where E_n is the nth note of the melody. The probability for an entire melody given the key then multiplies this quantity over all notes:

$$P(\text{melody} \mid \text{key}) = \prod_n P(E_n \mid \text{context}, \text{key}) \qquad (7)$$

Multiplying this by the prior probability of a key yields the joint probability of a melody and a key:

$$P(\text{melody}, \text{key}) = P(\text{key}) \prod_n P(E_n \mid \text{context}, \text{key}) \qquad (8)$$

(Choosing the key that maximizes this quantity yields the most probable key given the context—in effect, providing a key-finding algorithm [Temperley, 2007]; but this is not our concern here.) Summing this over all keys yields the overall probability of the melody:

$$P(\text{melody}) = \sum_{\text{key}} (P(\text{key}) \prod_n P(E_n \mid \text{context}, \text{key})) \qquad (9)$$

This can be used to calculate the probability of any sequence of pitches. It is also true that

$$P(E_n \mid \text{context}) = P(E_0 \ldots E_{\text{n}}) / P(E_0 \ldots E_{n-1}) \qquad (10)$$

Each of the two terms on the right (the context plus the current note, and the context without the current note) can be treated as a melody and its probability calculated using equation (9) shown above; the ratio between them is then the probability of the note given the context. In effect, this considers all possible keys in predicting the next note, giving more weight to keys that are more probable given the context.

A further issue concerns the kind of training data to use. Ideally one would want to use data that reflects the musical experience of the subjects, but this is difficult to know. (Cuddy and Lunney's [1995] experiment used one group of musician subjects and one group of non-musicians; the data used here averages the mean ratings from the two groups.) In previous work the Essen folksong corpus was found to yield good results, so that was initially used here.

The models were tested by calculating the correlation between each model's judgments and the expectedness ratings. Log probabilities rather than raw probabilities were used, since that yielded better results on earlier tests (Temperley, 2007). Certain aspects of the models were optimized on the data, as they were on the corpus tests. For the Markov model, different orders of the interval and scale-degree components were tried. The best combination was found to be a zeroth-order scale-degree model (as was found also for the corpus tests presented earlier) and a first-order interval model; this yielded a correlation of .79. (A second-order interval model was not possible, since each continuation tone was only preceded by two notes.) For the Gaussian model, the variances of the proximity and range profiles were optimized; this yielded a correlation of .81. The models were also tried with different corpora for training; out of the four corpora considered in the last section, the one yielding best performance for the Markov model was the classical corpus, with a correlation of .82. For the Gaussian model, no corpus yielded better performance than the folksong corpus.

In earlier work (Temperley, 2007, 2008) I presented a melodic expectation model very similar to the Gaussian model used here, but with some additional features added. The model calculates the mean of the range distribution in a complex way that takes into account both the mean of the pitches heard so far and a general preference for a certain absolute range. It also assigns major keys a somewhat higher prior probability than minor keys (since major tends to be more common), and it modifies the scale-degree profile for the last note of the melody, boosting the value for the tonic scale degree, on the grounds that listeners tend to interpret the last note of any melody as the tonic. (Cuddy & Lunney's [1995] and Schellenberg [1997]'s models also reflect this preference.) As noted earlier, adding these features to the Gaussian model improved its correlation to .88. It seemed possible that adding these features to the Markov model would yield a similar improvement. Adding a preference for major keys and a "last-note-as-tonic" preference improved performance only slightly, however, from .82 to .83. There seemed to be no simple

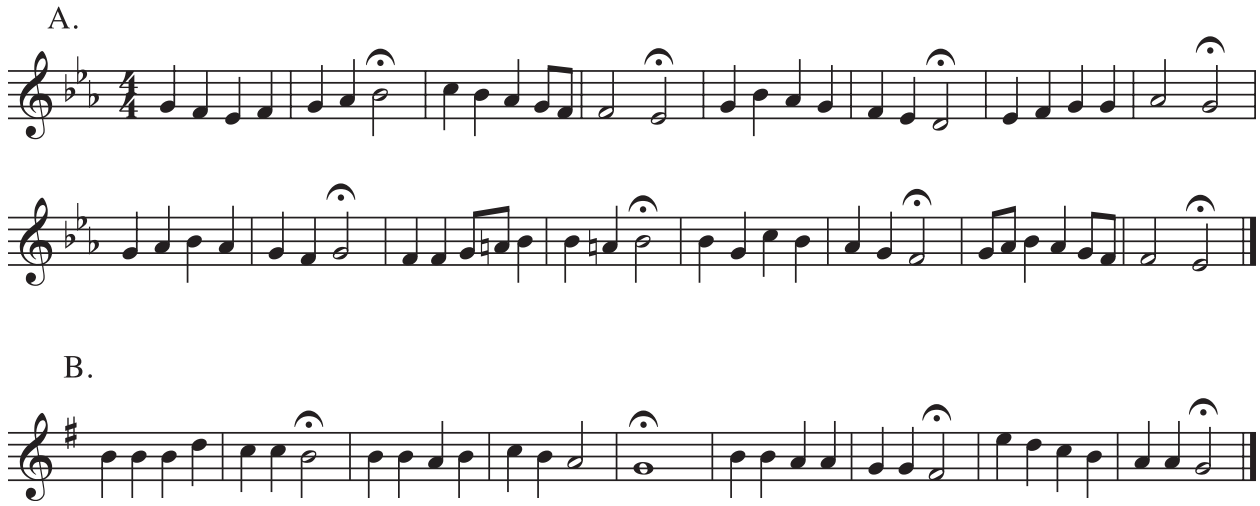

FIGURE 3. The two Bach chorale melodies used in Manzara et al.'s (1992) experiment: (A) BWV 159, (B) BWV 379.

way of incorporating the absolute range preference, so this was not attempted.

Cuddy and Lunney's (1995) study used very simple, one might say artificial, stimuli; one might wonder how the models would perform in more musically realistic situations. Suitable data sets are difficult to find. One widely used data set is from a study by Schellenberg (1996), in which subjects heard fragments of English folk songs and had to rate the expectedness of possible continuations. The problem with this data set is that only diatonic test tones were used as continuation tones; since the most salient tonal distinction between pitches is between diatonic and chromatic tones, the data set is of little interest for models that seek to capture effects of tonality.

Another (small) body of data is provided in a study by Manzara, Witten, and James (1992). This study used an ingenious experimental paradigm in which subjects (both musicians and nonmusicians) heard a portion of a Bach chorale melody and then bet pretend money on possible continuations; the amount bet on the actual note was taken as an indicator of its subjective probability. The procedure was used on two Bach chorale melodies, BWV 159 and BWV 379 (see Figure 3); the authors reported the log probability of each note, averaged across subjects. This data can be used in a similar manner to the Cuddy/Lunney data. In this case the models make a judgment about each note, based on the previous context. The first three notes of each melody were skipped, thus allowing the possibility of a second-order interval model; this left a total of 80 data points. While the key is indicated in the score of each chorale, it was not assumed that participants would have this information; key was treated as indeterminate, as with the Cuddy and Lunney (1995) data. For training data, it seemed most natural to use the chorale corpus used earlier. (BWV 379 was originally part of this corpus, but was removed from it for this test.) It is not obvious that participants in the study would have applied probabilities drawn from the chorale repertoire, or that they would even be familiar with this repertoire; but this seemed at least a reasonable starting point.

For the Gaussian model, optimizing the range and proximity variances yielded a correlation of .42. For the Markov model, the optimal configuration proved to be a first-order interval model combined with a zeroth-order scale-degree model, yielding a correlation of .52. The other three corpora used in earlier tests were also tried as training sets. None of them yielded any improvement for the Gaussian model; the classical corpus gave a very small improvement (.53) for the Markov model.

Overall, the tests presented here present a mixed picture as to the relative success of the Markov and Gaussian models in modeling expectation data. The Markov model's fit to the Cuddy and Lunney (1995) data is about the same as that of the Gaussian model; on the chorale data, the Markov model fares substantially better. Simplicity would appear to be a point in favor of the Gaussian model, though since the best Markov model employs only first-order interval probabilities, the difference in complexity is considerably less than on the earlier corpus tests. As noted earlier, other models employing a Gaussian approach to interval (Temperley, 2007) and a Markovian approach (Pearce & Wiggins, 2006) have been proposed that exceed the performance

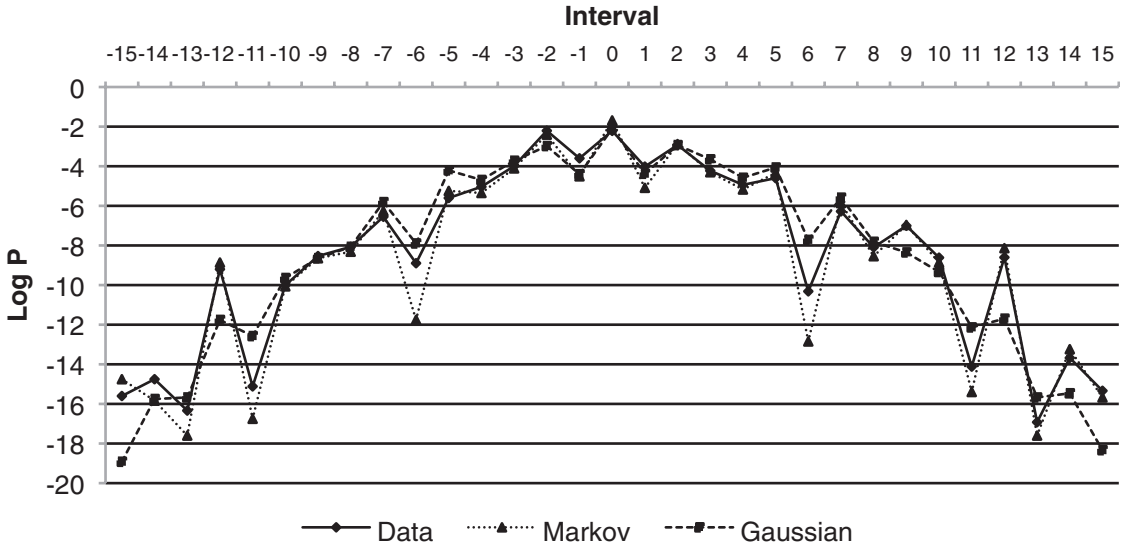

FIGURE 4. Three distributions of melodic intervals: from the Essen Folksong Collection, and from melodic sequences (one million notes) generated by the Markov model and Gaussian model. Probabilities are shown in log base 2.

of the simple models presented here, but the many differences between them make comparison difficult. (Witten, Manzara, & Conklin [1994] and Pearce & Wiggins [2006] also tested their multiple-viewpoint models on the chorale expectation data; Pearce & Wiggins report a correlation of .80.) The aim of the current experiment was to provide a more controlled comparison of the two approaches.

MODELING THE DISTRIBUTION OF MELODIC INTERVALS

As a final test, we return to an issue discussed in the first section of this paper. I posed the question: which of the two models presented here can more successfully predict the distribution of intervals in the Essen folksong corpus, shown in Figure 1 (and again in Figure 4)? We can address this question by using the two models in a generative fashion. Given a starting point, each model creates a probability distribution for the next note; this distribution can be sampled, choosing a note at random. Repeating this process and adding each note to the context for the next note, we can create a long note sequence and examine the resulting interval distribution. Using the Gaussian model—considering scale-degree probabilities (extracted from the corpus), proximity to the preceding note, and proximity to the mean position of the preceding pitches—a melody of 1 million notes was created; the resulting interval distribution is shown in Figure 4. Kullbach-Leibler (K-L) divergence, a standard measure of the similarity between two probability distributions, yields a value of 0.08 between this and the observed interval distribution of the corpus (a lower value indicates a better fit between the distributions). Qualitatively, it can be seen that the model fits the data well, capturing the large-scale "bell" shape of the distribution as well as many of the local peaks and valleys. The range profile has little effect here; almost the same K-L divergence, 0.07, was obtained by a model that considered only proximity and scale degree.

For the Markov model, we can ask, which version—that is, which orders of interval and scale degree—yields the best fit with the interval distribution of the corpus? The answer is simple. Since the data in Figure 1 represents the zeroth-order frequency of intervals in the corpus, a melody produced by a zeroth-order interval model trained on this same distribution—without considering scale degree—will reproduce it perfectly (or rather, will converge on this distribution as the length of the melody approaches infinity). This was verified empirically: The K-L divergence between the observed interval distribution and that of a 1-million-note melody generated from a zeroth-order interval model was less than 0.0001; the same was true of first-order and second-order interval models. Since the distribution produced by a purely interval-based Markov model is already perfect, factoring in scale degree can only worsen it. And indeed, multiplying the interval probabilities by scale-degree probabilities, as was done for the Gaussian model on the previous test, increases the K-L divergence, yielding 0.10 for a zeroth-order interval model, 0.07 for a first-order interval model, and 0.05

for a second-order interval model. The distribution for the second-order interval, zeroth-order scale-degree model—the best version of the Markov model in the corpus tests presented earlier—is shown in Figure 4.

To understand why the Markov model behaves as it does, consider again the case of whole steps versus half steps. A Markovian interval model, with no influence of scale degree, will capture the preference for whole steps over half steps by exactly the right amount. But as noted earlier, this preference may also be seen as arising from scale-degree probabilities: moving from one diatonic scale degree to another is more likely to involve whole step than half-step motion. Thus, multiplying scale-degree probabilities by interval probabilities makes the difference between whole steps and half steps greater than it should be; this can be seen from Figure 4, where the Markov model underestimates the probability of half steps. In other respects, the Markov model is able to capture features of the interval distribution that the Gaussian model misses, such as the preference for descending versus ascending steps (seen in Figure 4); and overall, the K-L divergence for the second-order Markov model (with scale-degree probabilities) is lower than that of the Gaussian model. We should remember, also, that incorporating scale-degree probabilities *improves* the Markov model with regard to the modeling of sequential melodic data (as shown in Table 1). Still, the issue raised here suggests that multiplying interval and scale-degree probabilities may not be the ideal solution.

Discussion

The current study compared two probabilistic methods of modeling melodic interval. Under the Markov method, the probability of an interval is defined by its count in a corpus, conditioned on previous intervals. Under the Gaussian method, it is a simple function of the size of the interval to the previous note and the distance to the mean pitch of the melody. In both models, this interval probability was then multiplied with the probability of the scale degree of the note. The models were tested on their ability to predict three kinds of data: sequential data from melodic corpora, experimental data from melodic expectation studies, and the distribution of melodic intervals in a corpus. Regarding the sequential corpus data, the Markov model yielded a somewhat better fit to the data than the Gaussian model on all four corpora examined. The difference in cross-entropy between the models ranged from 6% to 16%, depending on the corpus. The Markov model is, however, much more complex than the Gaussian model, requiring at least (very conservatively) several hundred times as many parameters. On expectation data, the Markov model's fit to the data was about the same as the Gaussian model on one corpus and considerably better on another; here again, simplicity would favor the Gaussian model. On the interval distribution, both models fit the data quite closely, the Markov model slightly beter than the Gaussian model; in this case, the best-performing version of the Markov model is one that does not consider scale degree at all.

As noted before, there is no widely accepted method for balancing simplicity against goodness-of-fit. The question is, again, which model is most plausible from a cognitive point of view. In terms of modeling the cognitive processes of composition: Does the better fit of the Markov model justify the view that composers maintain individual preferences for different intervals? Or, in view of the much greater simplicity of a Gaussian model, is it more that plausible that composers simply favor pitches that are close to the previous pitch and to the center of the range? Similar questions could be posed regarding the modeling of expectation. There appears to be no straightforward way of answering such questions. While one method for balancing simplicity and goodness-of-fit (the AIC) favors the Gaussian model, other measures might yield different results. Thus, I draw no conclusion here as to which model "wins" the current competition. I do maintain, however, that simplicity should be given *some* weight, so that a given difference in goodness-of-fit could potentially be outweighed if the difference in model complexity was great enough.

The multiple-viewpoint approach offers another way of combining scale degree and interval probabilistically, namely, through a *linked viewpoint* (Conklin & Witten, 1995). A linked viewpoint counts the occurrences of two features in combination, and defines probabilities accordingly. In the case of scale degree and interval, for example, the model would count the occurrences of scale degree 1 combined with a melodic interval of -2, and so on. This can also be implemented in a Markovian fashion, by counting the occurrences of longer sequences of feature combinations. The problem here is the explosion of parameters: If we limit the interval range to an octave (25 possible intervals) and allow only a single scale-degree profile, even a zeroth-order model of this kind would require $12 \times 25 = 300$ parameters; a first-order model would require 90,000; a second-order model would require 27 million. It seems unlikely that such a model would yield an improvement in cross-entropy that would justify this level of complexity, though the possibility might be worth exploring. The

same point might be made about higher-order (e.g., third-order) Markov models, which might improve goodness-of-fit but only at the expense of a huge increase in complexity. Another possibility would be to adopt a "variable-order" model, in which the order used by the model can vary depending on the context (Pearce & Wiggins, 2004).

It might also be possible to modify the Gaussian model to improve its performance. One could maintain the factors of pitch proximity and range, but build in further principles to better fit the data, such as the fact that steps are more likely to be descending and skips are more likely to be ascending (Huron, 2006). Also of interest here is the concept of *inertia*, also known as *process* (Larson, 2004; Narmour, 1990): a melodic step (ascending or descending major or minor second) is highly likely to be followed by another step in the same direction. In the Essen corpus, for example, 43.1% of steps are followed by a same-direction step, but only 18.3% by a different-direction step. The Gaussian model, as currently defined, does not capture this phenomenon; the Markov model does capture it, assuming a first-order (or higher) model of interval. One could incorporate inertia into the Gaussian model with a special rule that boosted the probability of an ascending step following a previous ascending step (and similarly for descending steps). This might well improve the performance of the Gaussian model, while avoiding the large number of parameters required by a fully Markovian approach.

Other factors influencing the structure of melodies might well be incorporated into the models presented here. One important factor is the repetition of patterns, such motives and themes. The multiple-viewpoint approach (Conklin & Witten, 1995) offers a solution to this problem by combining a "long-term" model, embodying general knowledge about the style (such as the kind of knowledge discussed in this paper), with a "short-term" model (also Markovian) trained only on previous notes within the current melody. Another important factor is harmony. It is generally assumed that an unaccompanied melody has a harmonic structure, which can be inferred from the notes and may then constrain the prediction of further notes. For example, if a melody begins with scale degrees $\hat{1}$-$\hat{3}$, this seems to imply tonic harmony, perhaps making it likely that the next note will also be part of the same harmony, such as $\hat{5}$ (though not *every* note is part of the harmony; nonharmonic tones, such as passing tones, also occur). Finally, there is the whole issue of rhythm and meter. This is an important dimension of melody that might itself be predicted by a probabilistic model (Temperley, 2010). It might also be used to inform predictions about pitch—for example, if some scale degrees tend to be longer in duration or to fall on stronger metric positions than others (an approach that has been used in some multiple-viewpoint studies, e.g., Conklin & Witten, 1995).

A premise of this study is that models of the kind presented here could conceivably represent part of the cognitive process of composing (or forming expectations for) tonal melodies. This assumption deserves some further clarification. I do not wish to imply that the creation of melodies is literally probabilistic or stochastic—that is, involving random choices. Rather, the constraints on interval and scale degree presumably interact with a variety of other constraints and preferences in some way to produce the final product. It is the nature of these constraints—in particular, intervallic constraints—that is at issue here: whether they take the form of fine-grained preferences for specific intervals, or more general preferences for maximizing proximity to the previous pitch and the center of the range. It is also possible, of course, that both kinds of preferences play a cognitive role.

The two models presented here might be seen to represent two approaches to the modeling of music cognition. One approach is *statistical*—involving the gathering of large number of statistical parameters from data. The other is *rule-based*—building a model based on a few simple principles. The term "rule-based" is sometimes used to refer to non-probabilistic models, but it seems logical to allow that a probabilistic model can be rule-based—though the rules involved tend to be gradient rather than categorical in nature. The distinction is not always clear-cut, however. The Gaussian model presented here is rule-based in its handling of interval, but statistical in its handling of scale degree; one might say that it represents a combination of rule-based and statistical approaches. Still, it is more toward the "rule-based" end of the spectrum than the Markov model, which is purely statistical in character.

One might ask whether the statistical approach or the rule-based approach to melodic interval is more cognitively plausible *a priori*, given what we know about cognition in general. This question admits of no simple answer. Debates over the relative merits of rule-based versus statistical approaches have arisen repeatedly in cognitive science, most famously perhaps in linguistics (Marcus, Vijayan, Bandi Rao, & Vishton, 1999; Pinker & Prince, 1988; Saffran, Aslin, & Newport, 1996); the issue is far from resolved. Both the Markov and Gaussian approaches to melodic interval have connections with proposals in other domains of

cognitive science. Markovian methods are widely used in psycholinguistics; for example, McDonald and Shillcock (2003) propose a bigram model to predict word reading times. The use of Gaussian functions is also well-established in cognitive modeling, in areas such as vision (Marr, 1982) and categorization (Shi, Griffiths, Feldman, & Sanborn, 2010). Clearly, humans have the capacity to learn general rules; they also have the capacity to absorb large amounts of statistical information. The current study certainly does not resolve this debate, though it perhaps adds one more piece of evidence to a very complicated picture. With regard to the modeling of musical interval, at least, the current study suggests that both statistical and rule-based approaches have points in their favor and deserve serious consideration.

Author Note

I am grateful to David Huron for making available to me his encoding of the Barlow & Morgenstern corpus, to Leonard Manzara for giving me permission to use the data from Manzara, Witten, and James (1992), and to Marcus Pearce for sending me that data.

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

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716-723.

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

Barlow, H., & Morgenstern, S. (1948). *A dictionary of musical themes.* New York: Crown Publishers.

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

Chai, W., & Vercoe, B. (2001). Folk music classification using hidden Markov models. *Proceedings of the International Conference on Artificial Intelligence.*

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

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

Deutsch, D. (1999). Grouping mechanisms in music. In D. Deutsch (Ed.), *The psychology of music* (2nd ed., pp. 349-411). San Diego, CA: Academic Press.

Gauldin, R. (1985). *A practical approach to sixteenth-century counterpoint.* Englewood Cliffs, NJ: Prentice-Hall.

Hartmanis, J., & Stearns, R. (1965). On the computational complexity of algorithms. *Transactions of the American Mathematical Society, 117*, 285-306.

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. Cambridge, MA: MIT Press.

Klapuri, A., & Davy, M. (2006). *Signal processing methods for music transcription*. New York: Springer.

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

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

Marcus, G.F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*, 77-80.

Marr, D. (1982). *Vision*. San Francisco, CA: Freeman.

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

McDonald, S., & Shillcock, R. (2003). Low-level predictive inference in reading: the influence of transitional probabilities on eye movements. *Vision Research, 43*, 1735-1751.

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

Pearce, M., Conklin, D., & Wiggins, G. (2005). Methods for combining statistical models of music. In U. Wiil (Ed.), *Computer music modelling and retrieval* (pp. 295-312). Berlin: Springer.

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*, 377-405.

Pearce, M., & Wiggins, G. (2012). Auditory expectation: The information dynamics of music perception and cognition. *Topics in Cognitive Science, 4*, 625-652.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28*, 73-193.

Pitt, M., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Journal of Experimental Psychology (General)*, *109*, 472-491.

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

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry.* Hackensack, NJ: World Scientific Publishing Company.

Rolling Stone (2004). The 500 greatest songs of all time. *Rolling Stone*, *963*, 65-165.

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

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928.

Schaffrath, H. (1995). *The Essen folksong collection.* D. Huron (Ed.). Stanford, 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.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Shi, L., Griffiths, T., Feldman, N., & Sanborn, A. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin and Review, 17*, 443-464.

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

Temperley, D. (2008). A probabilistic model of melody perception. *Cognitive Science, 32,* 418-444.

Temperley, D. (2010). Modeling common-practice rhythm. *Music Perception*, *27*, 355-376.

Temperley, D. (2012). Computational models of music cognition. In D. Deutsch (Ed.), *The psychology of music* (3rd ed., pp. 327-368). Amsterdam: Elsevier.

Temperley D., & de Clercq, T. (2013). Statistical analysis of harmony and melody in rock music. *Journal of New Music Research, 42*, 187-204.

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.

Witten, I., Manzara, L., & Conklin, D. (1994). Comparing human and computational models of music prediction. *Computer Music Journal, 18(1)*, 70-80.