Modeling Common-Practice Rhythm

David Temperley
Eastman School of Music

This study explores ways of modeling the compositional processes involved in common-practice rhythm (as represented by European classical music and folk music). Six probabilistic models of rhythm were evaluated using the method of cross-entropy: according to this method, the best model is the one that assigns the highest probability to the data. Two corpora were used: a corpus of European folk songs (the Essen Folksong Collection) and a corpus of Mozart and Haydn string quartets. The model achieving lowest cross-entropy was the *First-Order Metrical Duration Model*, which chooses a metrical position for each note conditional on the position of the previous note. Second best was the *Hierarchical Position Model*, which decides at each beat whether or not to generate a note there, conditional on the note status of neighboring strong beats (i.e., whether or not they contain notes).When complexity (number of parameters) is also considered, it is argued that the Hierarchical Position Model is preferable overall.

Received March 2, 2009, accepted January 7, 2010.

Key words: rhythm, meter, music composition, cross-entropy, probabilistic modeling

Figure 1 shows two rhythmic patterns. one is from a classical-period piece; the other is a hypothetical pattern, not from any piece. It will probably be clear that the first of the two patterns is the classical one (in fact, it is from the opening melody of the third movement of Mozart's piano sonata K. 333; see Figure 3). The fact that we are able to distinguish a classical rhythm from a non-classical one suggests that classical rhythms are characterized by general principles; and it seems reasonable to suppose that these principles were operating, in some form, in the minds of classical composers. This raises the question: what are these general principles? That is, what is (was) the nature of the musical knowledge that led classical composers to write certain rhythmic patterns and not others?

Figure 2 shows a second pair of rhythmic patterns. One is from a classical piece; the other is from a European folk song. In this case, identifying the classical rhythm is probably more difficult (in fact, it is the second one; see Figure 3). This illustrates a second point: The rhythmic practice of the classical period has much in common with that of a range of other musical styles: Western art music from earlier (Baroque) and later (Romantic) periods, as well as much pre-twentieth-century European folk and popular music. This is not to say that the rhythmic practices of these styles are identical—one can often distinguish a Bach rhythm from a Brahms rhythm, and either of these from the rhythm of a folk song—but rather that there are certain fundamental principles that they all share. One could, indeed, speak of a rhythmic "common practice" in European music of (roughly) the seventeenth through nineteenth centuries, analogous to the well-known harmonic "common practice" that characterizes roughly the same body of music. My aim in the current study is to elucidate the principles underlying this rhythmic common practice.

If we define the field of music cognition broadly as the scientific study of the mental processes and representations involved in all kinds of musical experiences and behaviors, then studying the cognitive processes involved in composition is an entirely appropriate goal for the field. The pursuit of this goal raises formidable problems, however. To study compositional processes using experimental methods—the most common methodology of music cognition—is often difficult if not impossible. In general, hypotheses about creative processes do not easily lend themselves to experimental testing. In addition, much of the music that is of interest to us was written hundreds of years ago, and few possessors of this compositional expertise are available today. However, we may still test claims about composition using the results of these compositional processes—the music itself. Music provides a body of data (albeit not experimentally controlled data) that we can seek to model, just as we would any other data; the model that makes the most accurate predictions about the data is then the most plausible model of the cognitive processes that gave rise to it. In modeling such data, we try to construct models that generate (or in some other way predict) patterns that were

Music Perception volume 27, issue 5, pp. 355–376, issn 0730-7829, electronic issn 1533-8312 DOI:10.1525/mp.2010.27.5.355

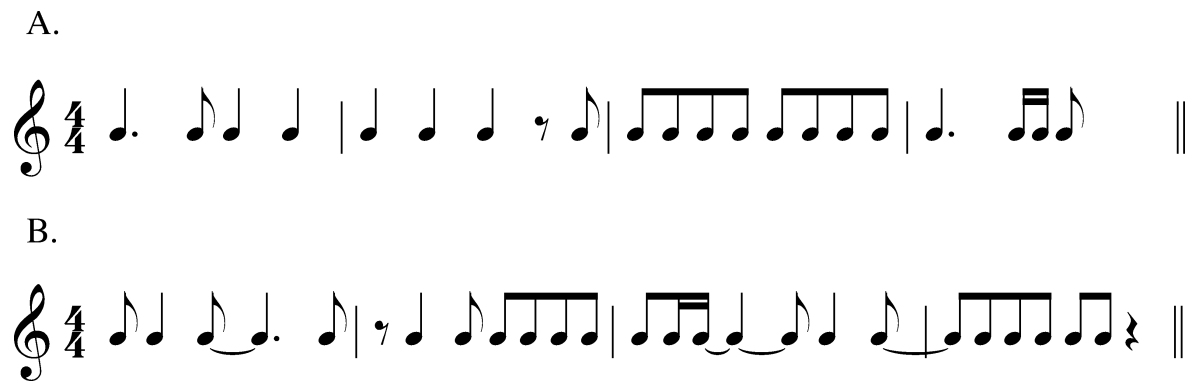


FIGURE 1. Two melodic rhythms.

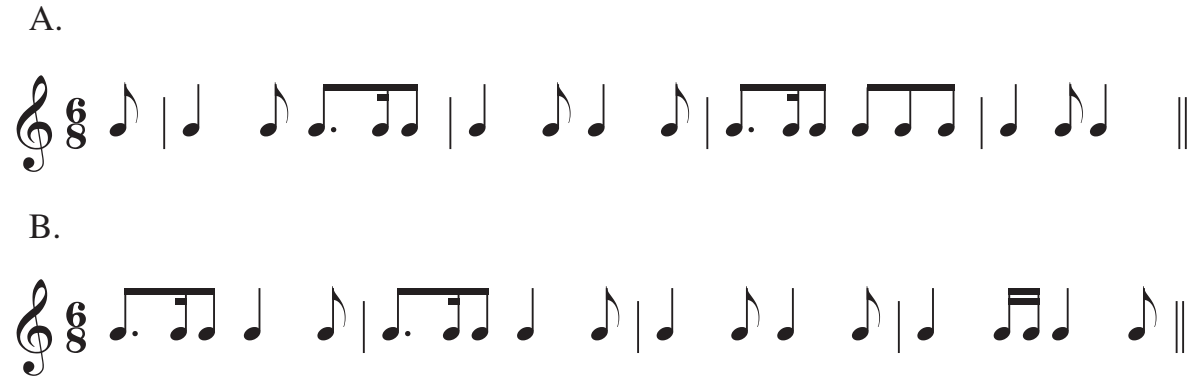


FIGURE 2. Two more melodic rhythms.

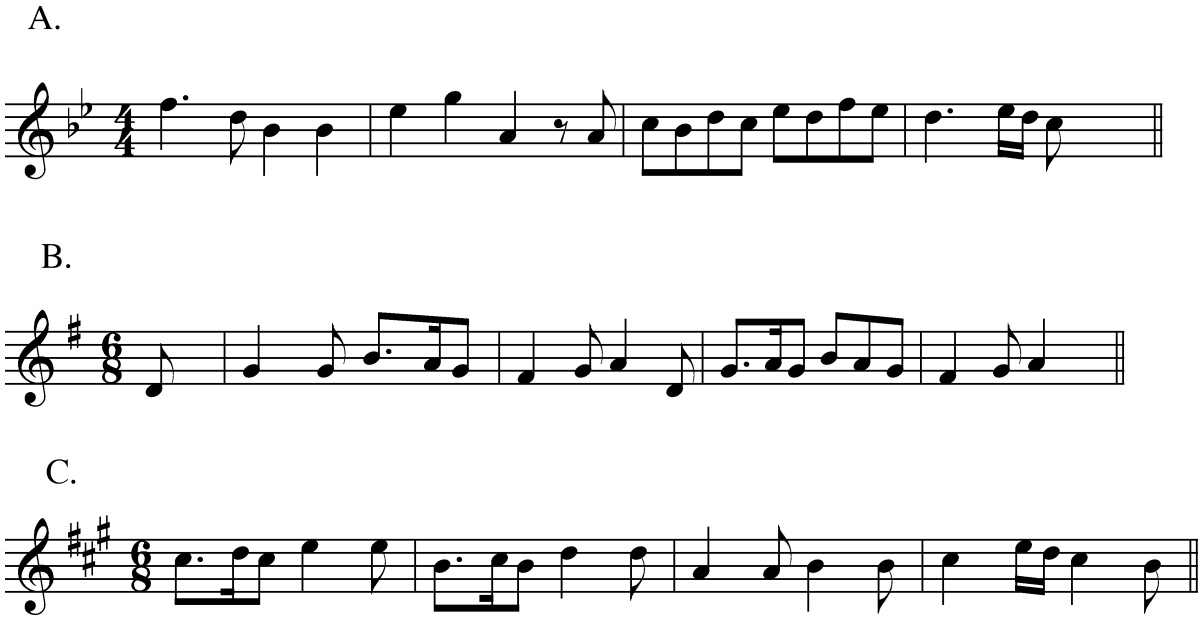


FIGURE 3. (A) The melody of Figure 1A: Mozart, Sonata K. 333, third movement, mm. 1–4. (B) The melody of Figure 2A: "Verschlafener Jaeger es wollt ein Jaegerli jagen," from the Essen Folksong Collection. (C) The melody of Figure 2B: Mozart, Sonata K. 331, first movement, mm. 1–4.

actually written by common-practice composers (such as Figure 1A above), and fail to generate those that were not (such as Figure 1B). A model's success at this task is one criterion—not the only criterion, but certainly an important one—whereby we might evaluate it as a characterization of the cognitive processes underlying the creation of common-practice rhythms.

Perhaps the most impressive attempt to model compositional data in recent years has been the work of Huron (2001, 2006). Much of Huron's work in this area has been concerned with the ways that composition is shaped by principles of auditory perception. Huron uses this reasoning, in the first place, to propose principled explanations of well-known rules of composition. For example, it is desirable from a compositional viewpoint for the independent lines of a polyphonic piece to remain perceptually distinct. Perfect consonances (perfect fifths and octaves) cause simultaneous notes to fuse, as does commodulation (two voices moving simultaneously by the same interval). Thus, the prediction is that composers should tend to avoid commodulating perfect consonances, also known as parallel fifths and octaves; and indeed, a traditional contrapuntal rule prohibits such motions (Huron, 2001). Huron also uses this approach to generate new predictions that are not covered by traditional rules, but prove to be borne out in studies of musical corpora. For example, changes of texture involving the departure of a single line appear to be less easily perceived than additions of a single line; composers seem to have responded to this tendency by avoiding single-line departures, preferring to retire several voices from the texture at once (Huron, 1990).

A very different approach to explaining compositional practice is reflected in the work of Cope (2005) and Gjerdingen (2007). According to these authors, composition largely entails the reproduction and concatenation of patterns that have been heard in other music and are stored in memory. For Cope, the patterns involved are literal configurations of notes; for Gjerdingen, they are more abstract "schemata," skeletal scale-degree patterns that may be elaborated in an endless variety of ways. As an example of this "replicative" approach, let us return to Mozart's rhythmic pattern in Figure 1A. One might observe that this rhythm can be broken down into four one-measure patterns, each of which is quite characteristic of the classical style and could undoubtedly be found in innumerable other classical-period pieces. By contrast, the one-measure units of Figure 1B are much less characteristic of classical pieces; the reason Mozart never wrote such patterns, one might suggest, is that he never heard them in the music of the time and thus never incorporated them into his musical vocabulary.

The replicative view of composition is in some ways quite compelling. There seems to be no doubt that composers (like other listeners) store in memory musical patterns that they hear frequently and sometimes reproduce these patterns in their compositions. However, this view also has limitations. In particular, it has difficulty explaining the fact that the music of the classical period (or indeed any other style) seems to adhere to certain basic, consistent principles. For example, to return to an example mentioned previously, Mozart tends to avoid parallel fifths and octaves. A replicative account could only explain this by saying that the music Mozart heard tends to avoid parallel fifths and octaves; but this merely defers the question, since it must then be explained why the music Mozart heard had this consistent property. A similar point could be made about rhythm. In the case of Figure 1A, for example, it can be seen that the notes in Mozart's rhythm have a strong tendency to occur on relatively strong beats of the meter; 12 of the 20 notes in this rhythm occur on strong eighth-note beats, whereas only 6 of the 20 notes in Figure 1B occur on strong eighth-note beats. And this reflects a general fact about common-practice rhythm (as we will see below). An account of composition that views pieces simply as concatenations of patterns drawn from memory does not seem to offer any explanation for such regularities. (Here I assume the well-known view of metrical structure as a framework of levels of beats of varying strength, as shown above the staff in Figure 4.

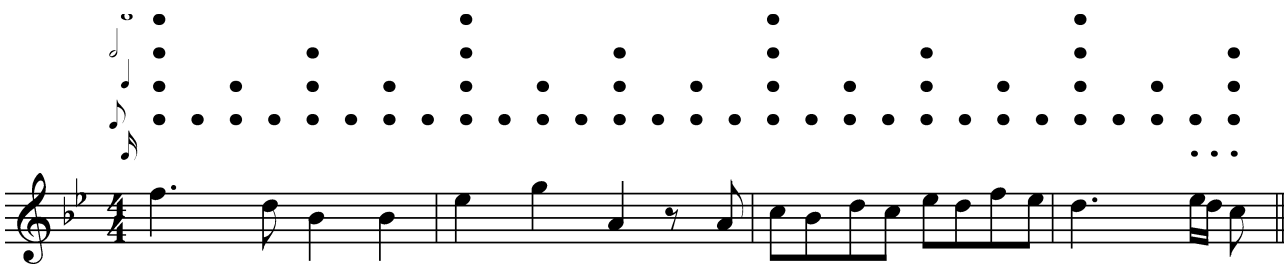


FIGURE 4. The melody in Figure 3A, showing metrical grid.

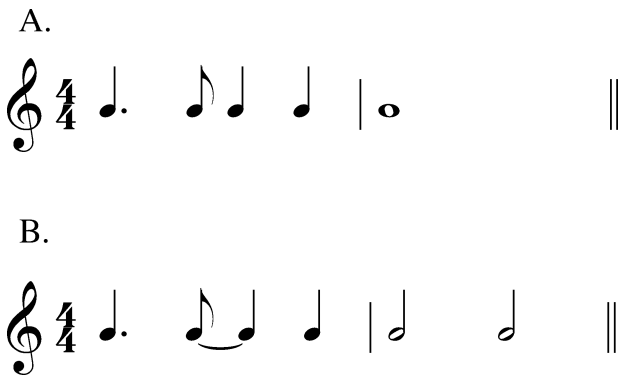


FIGURE 5. Two patterns with the same distribution of notes on beats.

Each level corresponds to a rhythmic value; a "strong" eighth-note beat is one that is present at one or more levels above the eighth-note level, while a "weak" eighth-note beat is present only at the eighth-note level.)

Thus, it is difficult to escape the conclusion that the creation of common-practice rhythms involved general principles of some kind. But the exact nature of these principles is unclear. Consider the principle, stated above, that notes are more likely to occur on strong beats; this is the essence of a simple model that I present below, which I call the *Metrical Position Model*. While this model captures an important regularity, it is imperfect as a characterization of classical-period rhythmic practice. According to this model, the rhythm in Figure 5A is just as likely as that in Figure 5B; both of them have the same distribution of notes on beats (two notes on whole-note beats, one note on a half-note beat, one note on a quarter-note beat, and one note on an eighth-note beat). Yet Figure 5A is plainly more characteristic of common-practice rhythm than Figure 5B. Clearly, then, the Metrical Position Model is inadequate as a model of common-practice rhythm. The question then arises, what other kinds of principles might better capture the facts?

A solution to this problem arises from another observation about Figure 1A: among the notes that occur on weak beats (at the quarter-note level or below), all are adjacent to notes on stronger beats. For example, the one note on a weak sixteenth-note beat (in the fourth measure) is flanked by notes on both of the immediately adjacent eighth-note beats; every note on a weak eighth-note beat has notes on at least one of the adjacent quarter-note beats; and similarly for notes on weak quarter-note beats. Similarly, every weak-beat note in Figure 5A is adjacent to a strong beat with a note; in Figure 5B, however, this is not the case (the second note has no adjacent strong-beat note). Perhaps, then, common-practice rhythms are generated in a hierarchical manner: notes on strong beats are generated first, and notes on weak beats are then conditional on the adjacent stronger ones. This idea is the basis for a model of rhythm that I will call the *Hierarchical Position Model*.

In light of cases such as Figure 5, it may seem likely that the Hierarchical Position Model will predict common-practice rhythms better than the Metrical Position Model. But our intuitions about such matters are notoriously unreliable; what is needed is a rigorous, objective way of determining which model fits the data better, and by how much. In what follows, I explore a quantitative method for testing models of rhythm on musical corpora. The models will be tested both on classical-period art music (as represented by Haydn and Mozart string quartets) and European folk songs. Our method of testing the models will be probabilistic. A model can be tested as to the probability it assigns to a body of data; the higher the probability, the better the model. As well as the two models described above, four other models also will be examined using the same method of evaluation.

It seems natural to suppose that the basic principles underlying composition—whatever they may be—play a role in perception as well. It is presumably these principles, at least in part, that allow listeners to judge whether a piece of music is characteristic of a style or not—an ability that most listeners have, at least to some extent. (I appealed to this ability earlier in this article, in asking the reader to decide which of the two rhythms in Figure 1 was taken from a classical-period piece.) After evaluating our six models with regard to compositional practice, I will consider their plausibility with regard to the perception of rhythm, and will consider several sources of evidence in this regard.

Testing Six Models of Rhythm

THE PROBABILISTIC METHOD OF MODELING COMPOSITIONAL PRACTICE

The idea that we can evaluate models of data by the probability that they assign to the data rests on firm, and quite simple, mathematical reasoning. Let us suppose we are given a body of data D and want to find the best model, M, of the source that gave rise to the data. In probabilistic terms, we want to find the most likely source model given the data, or the M that maximizes $P(M \mid D)$. Bayes' Rule, a basic rule of probability, tells us that for any M and D,

$$P(M \mid D) \propto (P(D \mid M)\, P(M) \qquad (1)$$

where $P(D \mid M)$ is the probability of the data given the model (known as the *likelihood* of the data), $P(M)$ is the probability of the model itself before the data are seen

(known as the *prior probability* of the model), and "∝" means "proportional to." If we assume that all models are equal in prior probability, then

$$P(M \mid D) \propto (P(D \mid M) \tag{2}$$

—that is, the most probable model given the data is simply the one that assigns highest probability to the data. The conditional clause in the previous sentence deserves emphasis: expression (2) only holds if the models under consideration are equal in prior probability. In some cases, there may be other considerations—for example, historical evidence about how composers thought about rhythm, neurological or experimental evidence about general cognitive mechanisms, or considerations of simplicity or parsimony—that lead us to assign higher prior probability to some models than others. But for now, let us neglect such factors and assume that all models are equal in prior probability.

The method of model evaluation just described is very well-established in cognitive science and machine learning. It is sometimes known as "maximum likelihood estimation"; mathematically it hinges on the concept of "cross-entropy," which will be discussed further below.[1] Perhaps the most well-known use of this technique is in speech recognition (Jurafsky & Martin, 2000). Probabilistic models of speech recognition require an estimate of the prior probability of any sequence of words; for this purpose, some kind of model of the language, or "language model," is needed. Language models can be evaluated by the probability they assign to a corpus of sentences; the higher the probability, the better the model. A very common kind of language model is a *Markov model*, which calculates the probability of each word conditional on the previous N words. If $N = 1$, it is a "bigram" model, in that it looks at the probabilities of pairs of adjacent words; if $N = 2$, it is a "trigram" model. It is important to emphasize that language models in computational linguistics generally are used simply for the practical purpose of speech recognition, and are not claimed to represent human language production. Here we extend the cross-entropy approach further, using it to actually evaluate models of the generative process.

Most often, the body of data we wish to model is a sample of items drawn from a larger population (perhaps a theoretically infinite population, such as the sentences of a language). The model must assign a probability to each item, $P(I)$; the probability assigned to the data as a whole is then the product of the probabilities for all the items,

$$P(D \mid M) = P(I_0) \times (P(I_1) \ldots \times (P(I_n) \tag{3}$$

To avoid the tiny numbers that result from multiplying many probabilities together, we can take the log of this expression (which does not change the results in terms of the ranking of different models):

$$\begin{aligned} \log P(D \mid M) &= \log(P(I_0)) + \log(P(I_1)) \ldots \\ &\quad + \log(P(I_n)) \\ &= \Sigma_n \log(P(I_n)) \end{aligned} \tag{4}$$

If we divide this expression by the number of items N, we get a "per-item" measure of the probability assigned to the data by the model. We also add a negative sign to cancel out the negative sign that results from taking the log of a probability.

$$-\log P(D \mid M) \text{ per item} = -1/N \, \Sigma_n \log(P(I_n)) \tag{5}$$

This is essentially equivalent to the definition of cross-entropy.[2] Note that cross-entropy is always positive, and that *lower* cross-entropy implies a *higher* probability assigned by the model.

One potential problem in probabilistic model evaluation is "overfitting." Suppose we were given a corpus of 100 melodic rhythms and wished to find the model assigning it highest probability. A trivial approach would be to define a model that assigned a probability of 1/100 to each of the melodies in the corpus, yielding a (per-song) cross-entropy of $-\log(1/100) = 4.6$, which is, in fact, the best (lowest) that could be achieved for that corpus. But this model would not be very useful. It would be highly complex; it would also have no ability to generalize to unseen melodies. (Since the entire probability mass of 1 is used up by the corpus, any other melody would be assigned a probability of 0.) To foil such trivial "cheating" solutions, it is usually stipulated that the corpus on which models will be tested may not be seen in designing the models. The typical approach is to use part of the data for training the model (e.g., setting the parameters) and another part for testing, a technique known as "cross-validation." This is the approach used here.

[1] The term "maximum likelihood estimation" is normally applied to the process of choosing parameters for a single model, rather than choosing between different models, as we will do here.

[2] The cross-entropy of a model with a body of data is normally defined as $-\Sigma_x P(x) \log(P_m(x))$. This assumes a body of data consisting of a series of items x, where each x may occur many times; the contribution of each x to the cross-entropy is the probability assigned to x by the model, $\log(P_m(x))$, weighted by the count of x in the data (as a proportion of the total), $P(x)$. But in the current case, each of the N melodic rhythms is assumed to occur only once, thus $-\Sigma_x P(x) \log(P_m(x)) = -\Sigma_n 1/N \log(P(I_n))$, which is equivalent to the definition of cross-entropy in expression (4). The term "perplexity" is also sometimes seen; the perplexity between a model and data is just 2 to the power of the cross-entropy.

Cross-entropy has not been widely used in music research. A number of studies from the 1950's and 1960's used entropy—essentially, the cross-entropy of a data set with itself—as a way of measuring the complexity of styles and pieces (for a review of this work see Cohen, 1962); but these studies made no use of cross-entropy as a method of model selection. Perhaps the earliest musical application of cross-entropy was the work of Conklin and Witten (1995). Conklin and Witten focused on the problem of modeling pitch patterns. A pitch sequence can be represented in various ways: as a series of pitches, scale-degrees, melodic intervals, scale-degrees combined with metrical positions, and so on. Each of these types of data can be represented as a Markov chain (what they call a "viewpoint"), and each type of Markov chain assigns a probability to the pitch sequence; viewpoints also can be combined. Conklin and Witten compared the predictive power of different "multiple viewpoint systems" with regard to pitch patterns in Bach chorale melodies (see also Pearce & Wiggins, 2004). While Conklin and Witten's stated goal was to generate new music rather than to model compositional processes, their method is fundamentally similar to what is proposed here.

We use the approach outlined above to compare six different models of melodic rhythm. The models are tested first on a corpus of European folk songs—the Essen Folksong Collection—and second, on a corpus consisting of the first violin parts of Mozart and Haydn string quartets. We begin with the folksong corpus for two reasons. First, it provides a very large body of data and thus offers a better opportunity for choosing between alternative models. Second, our focus in this study is on the rhythm of melodies. While it seems safe to say that the Essen collection consists entirely of melodies, classical string quartets—even the first violin parts—do not; at some points in classical quartets, the first violin plays passagework material not normally considered melodic, or accompaniment patterns supporting a melody in another voice. Still, classical quartets offer an interesting corpus for analysis and a useful comparison to the folksong data.

TESTING THE MODELS ON THE ESSEN FOLKSONG COLLECTION

The Essen Folksong Collection (Schaffrath, 1995) contains over 6,000 European folk melodies, transcribed with metrical information (time signatures and barlines); the melodies were encoded by Huron (1999) in *kern* notation, a widely used format for computational music representation. To simplify the situation, we consider only melodies in 4/4 time; this yields a set of 1,585 melodies. (We will consider later how the models might be extended to other time signatures.) We also exclude songs with notes on beats below the eighth-note level; 350 songs were excluded for this reason. This yields a corpus of 1,235 melodies, containing a total of 13,786 measures. From this corpus, 247 melodies were selected randomly as a test corpus and the remaining 988 were used as a training corpus.

To further simplify testing, in each song, we consider only the portion from the first downbeat to the last downbeat, inclusive, thus omitting most of the last measure as well as any notes that precede the first downbeat. We disregard note-offsets—that is, we do not distinguish between (for example) "half-note" and "quarter-note plus quarter-rest." Thus, each rhythmic pattern is represented simply as a pattern of note-onsets. (These points also apply to our tests of the Haydn-Mozart corpus, to be discussed later on.) We assume a metrical grid consisting of eighth-note, quarter-note, half-note, and whole-note levels; the correct grid for each song can be inferred from the kern notation.

We begin with two very simple models and work upwards to the more complex models proposed in the previous section.

Model 1 (Uniform Position Model). A decision is made at each beat as to whether or not to generate a note. Note onsets are equally likely at all beats.

This model has just one parameter, which is the overall likelihood of a note occurring on a beat. From the training set, we find that 51% of all beats have note onsets. (The beats under consideration are those of the eighth-note level; recall that songs with notes at sub-eighth-note levels were excluded from the corpus. We use the eighth-note level even in songs that contain no eighth-notes.) Let us now define variables B_n, one for each beat in the test corpus; each B_n has the value "note" (if there is a note onset there) or "rest" (if there is not). The model assigns probabilities $P(B_n = \text{note}) = .51$ and $P(B_n = \text{rest}) = .49$ for all beats. The probability of a song rhythm is then the product of these values for all beats in the melody. For example, consider the opening of Figure 1A—the first measure plus the downbeat of the second measure, shown in Figure 6 (let us assume this is an entire song). Five of the nine beats—beats 0, 3, 4

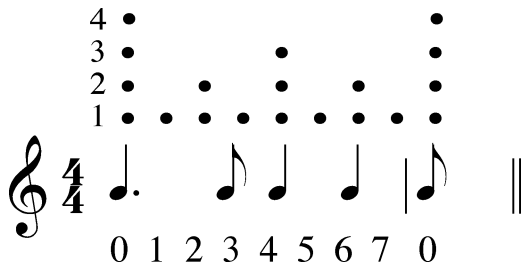


FIGURE 6. The beginning of the Mozart rhythm in Figure 1A.

TABLE 1. Cross-Entropy for Six Models of Rhythm on Two Corpora.

	Essen Corpus		HM Corpus	
	Cross-Entropy (per song)	Number of Parameters	Cross-Entropy (per piece)	Number of Parameters
1. Uniform Position Model	62.37	1	1150.14	1
2. Zeroth-Order Duration Model	54.45	15	943.89	15
3. Metrical Position Model	42.47	4	965.94	5
4. Fine-Grained Position Model	40.21	8	959.42	16
5. Hierarchical Model	38.76	13	819.37	17
6. First-Order Metrical Duration Model	37.36	56	783.79	240

and 6 of measure 1 and beat 0 of measure 2—have notes; the other four beats do not. (We label the eight beats of each measure as 0 through 7, with 0 being the downbeat.) Thus:

$$\begin{aligned} -\log\,(P(\text{rhythm})) &= -\log(.51 \times .49 \times .49 \times .51 \times .51 \\ &\quad \times .49 \times .51 \times .49 \times .51) \\ &= 6.22 \end{aligned} \tag{6}$$

In this way, we can compute the negative log probability for each song in the test set; adding these values (as in Equation 4 above) yields the negative log probability of the entire test set, which is 15,405. Recall, however, that cross-entropy normally represents negative log probability in a "per-item" fashion (as in Equation 5 above). In this case, it seems logical to treat songs as items; thus, we divide by the number of songs in the test set, yielding a (per-song) cross-entropy of 62.37 (see Table 1).[3] While we have nothing to compare this to at present, we will see shortly that this performance is very poor. This should not surprise us, for the model knows almost nothing; as far as it is concerned, every possible location for a note is as good as any other.

The model just presented operates by making a series of decisions about beats—that is, positions in time; it decides whether or not to generate a note-onset at each position. For this reason we call it a "position model." A very different approach to generating rhythms would be to make decisions for a series of notes, choosing the onset time for each note (under the assumption that each note must be later than the previous one). In effect, each decision determines the time interval between the previous note and the current one (usually known as an "interonset interval"). If we assume that each note extends to the onset of the following note (making no distinction between a note continuation and a rest), the interonset interval between two notes is equivalent to the duration of the first note; by choosing the location for one note, we choose the duration of the previous one. Thus we could call this a "duration model."[4]

Model 2 (Zeroth-Order Duration Model). A decision is made at each note as to its interonset interval from the previous note.

The term "zeroth-order" implies that each interonset interval is chosen independently of the previous one. We parameterize the model by gathering data as to the frequency of different interonset intervals in the training set (measuring intervals in eighth-note beats). These data are shown in Figure 7. The duration of a note could, in principle, be any value; but interonset intervals of more than two measures never occur either in the training

[3]It might seem more logical to treat beats as items, rather than songs. The problem with this is that some of the models presented below do not compute probabilities for beats, but rather, for notes. If these individual elements (notes or beats) were treated as items, then the number of items would differ between models, and it would not be possible to do a fair comparison between them.

[4]Both the position models presented here (Models 1, 3, 4, and 5) and the duration models (Models 2 and 6) are incomplete as generative models, in that they do not specify the length of the piece being generated. Position models could do this by generating a span of beats to be filled in with notes; duration models could do it by choosing a number of notes to be assigned metrical positions. In both cases, these choices could be made stochastically from distributions. However, to work out the details of this would be quite complex; I suspect also that it would have little effect on the results in terms of the relative cross-entropies assigned by the models to the test corpus. Model 2 is incomplete in another way: It does not assign any probability for the location of the first note of the melody. For that reason, the probability it assigns to the corpus is somewhat higher than it should be. Recall, however, that we exclude everything before the first downbeat. In virtually all melodies in the Essen corpus (all but two of the 988 melodies in the training set), the first note of the portion of the song considered is on the first downbeat (in other words, the first downbeat is almost never "empty"); if we assign this a probability of 1, the total probability assigned to the melody is unchanged. The same point applies to Model 6 below, which also does not assign any probability for the location of the first note.

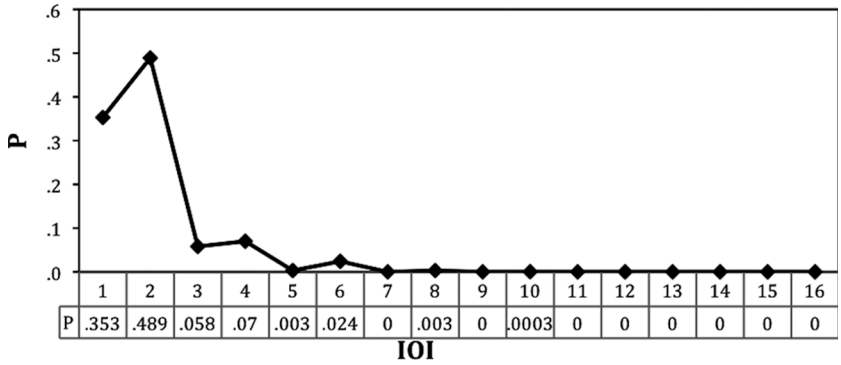


FIGURE 7. Interonset intervals (in eighth-notes) in the Essen training set.

corpus or in the test corpus; thus, they can be assigned a probability of zero. In testing, the probability of a rhythmic pattern is then the product of the probabilities for all of its durations. For the pattern in Figure 6 (with interonset intervals 3, 1, 2, 2):

$$\begin{aligned} -\log P(\text{rhythm}) &= -\log\,(.058 \times .353 \times .489 \times .489) \\ &= 5.32 \end{aligned} \tag{7}$$

For the corpus, the cross-entropy is 54.45 per song—somewhat better than the Uniform Position Model (see Table 1).

An essential flaw of both the Uniform Position Model and the Zeroth-Order Duration Model is that they have no knowledge of meter. As noted above, note-onsets are much more likely on strong beats of the meter; this is a well-known principle of music theory (Lerdahl & Jackendoff, 1983) and has been confirmed empirically as well (Huron, 2006; Palmer & Krumhansl, 1990; Temperley, 2007). Our next model captures this regularity.

Model 3 (Metrical Position Model). A decision is made at each beat whether or not to generate a note. The probability of a note at a beat depends on its metrical strength.

This is once again a position model, as it makes a decision for each temporal position. To set the model's parameters, we gather data as to the proportion of beats at each metrical level that have note onsets. The data are shown in Figure 8. Henceforth, level 1 (or L1) is the eighth-note level, level 2 is the quarter-note level, level 3 is the half-note level, and level 4 is the whole-note level. It can be seen, indeed, that the probability of a note onset increases monotonically with higher levels; the difference is greatest between the eighth-note and quarter-note levels and somewhat smaller for higher-level distinctions. As in Model 1, we define a variable B_n for each beat, but now $P(B_n)$ depends on the metrical level of the beat; for a level 1 beat, for example, $P(B_n = \text{note}) = .21$ and $P(B_n = \text{rest}) = 1 - .21 = .79$. The probability of a rhythmic pattern is again the product of the $P(B_n)$ values for all the beats. For the pattern in Example 6:

$$\begin{aligned} -\log P(\text{rhythm}) = -\log(&.988 \times .79 \times .311 \times .21 \times .842 \\ &\times .790 \times .689 \times .79 \times .988) \\ &= 4.00 \end{aligned} \tag{8}$$

(For example, the first beat is an L4 beat with a note, yielding a value of .988; the second beat is an L1 beat with no note, yielding .79; and so on.) The cross-entropy assigned to the corpus is 42.47 per song—a substantial improvement over our first two models.

While Model 3 captures a valid and important generalization about common-practice rhythms, it is inadequate in certain respects. A well-known principle of Western rhythm is that notes on strong beats tend to be longer than notes on weak beats (Lerdahl and Jackendoff, 1983). Given the rhythm "eighth-note/doubled-dotted-half," for example, it seems more natural to put the long note on the

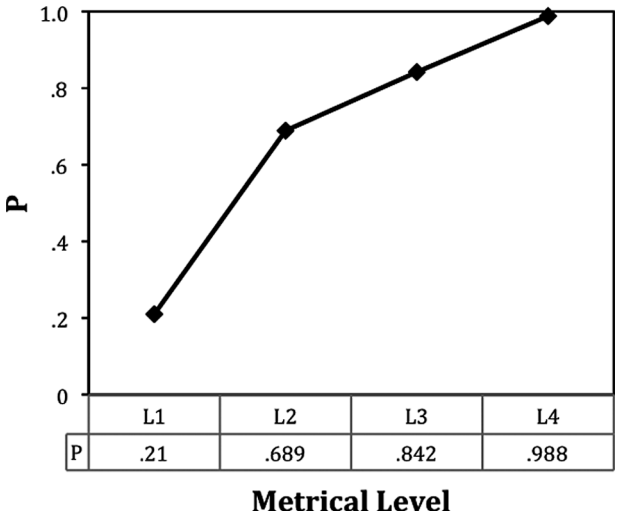


FIGURE 8. The proportion of beats with note onsets at different metrical levels in the Essen training set.

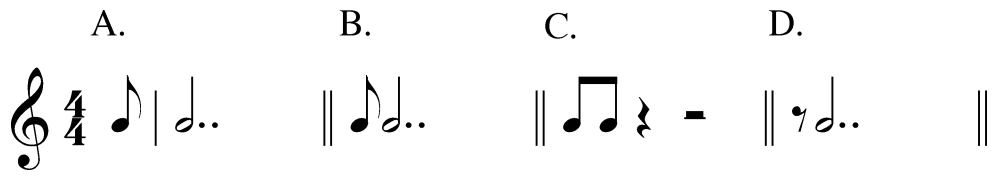


FIGURE 9. Four rhythmic patterns.

downbeat, as in Figure 9A, rather than the short note, as in Figure 9B. Model 3 has no way of capturing this; by this model, the two rhythms are assigned equal probability, since both have one note on a level 4 beat and one on a level 1 beat. One solution to this problem would be to condition the probability of notes on the position within the measure, as Model 3 does, but in a more fine-grained manner, distinguishing between different positions of equal strength. A note on a weak beat just before a strong beat (e.g., position 7) is likely to be short (since strong beats generally have notes), whereas a note just after a strong beat (e.g., position 1) is much more likely to be long. If we give higher probability to notes at position 7 than at position 1, we are in effect exerting pressure for weak-beat notes to be short. (This gives Figure 9A higher probability than Figure 9B, for example.) Thus, we define Model 4:

Model 4 (Fine-grained Position Model). A decision is made at each beat whether or not to generate a note. The probability of a note at a beat depends on its position within the measure.

In this case, separate parameters are set for the probability of a note at each of the eight positions within the measure. The data are shown in Figure 10. (It can be seen, as expected, that the probability for a note at position 7 is higher than for a note at position 1.) The cross-entropy calculations are then the same as in Models 1 and 3. The cross-entropy assigned to the corpus is 40.21 per song.

Our test results show that Model 4 predicts common-practice rhythms better than Model 3, but only slightly. This suggests that Model 4 may not be the best way of capturing the "note length" principle. Consider also the patterns in Figure 5, discussed earlier. These two patterns are identical by Model 3 and even by Model 4: the distribution of notes across beats of the measure is the same in both models (both patterns have two notes at position 0 and one note each at positions 3, 4, and 6). Yet Figure 5B clearly seems less characteristic of common-practice rhythm than Figure 5A. One might explain this in terms of the note length principle: Figure 5B features a long note at position 3 whereas Figure 5A does not. On the other hand, long notes on weak beats do sometimes occur. Patterns such as Figure 9B, while not common, are certainly not unheard of in common-practice music. (Here the literal durations of notes may make a difference; Figure 9C, in which the "long" note is a short note followed by a rest, seems a bit more idiomatic than Figure 9B, though by all the models considered here, the two are equivalent.) This suggests that perhaps "prefer long notes on strong beats" is not really the underlying principle

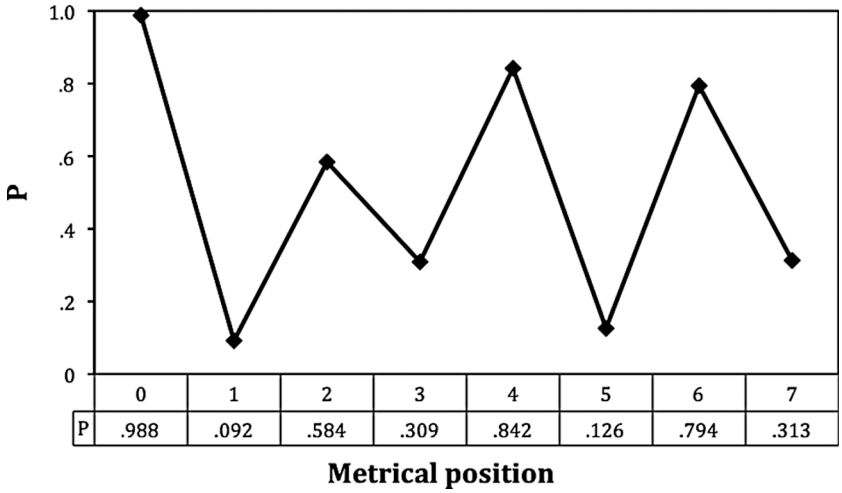


FIGURE 10. The proportion of beats with note onsets at each measure position in the Essen training set.

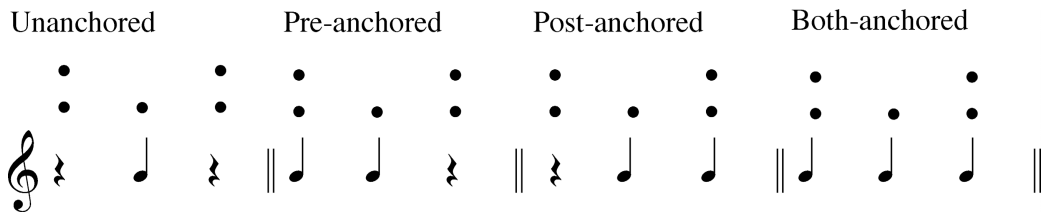


FIGURE 11. Context types in the Hierarchical Position Model.

involved. A further observation is that the weak-beat note in Figure 9B immediately *follows* a strong-beat note. By contrast, the weak-beat note in Figure 9D is not adjacent to a note on either side; and this pattern truly seems foreign to the common-practice idiom. Perhaps a note on a lower-level beat is more likely when there are notes on one or both of the adjacent strong beats: one might say in that case that the weak-beat note is "anchored" to the neighboring strong-beat note(s). (This term is due to Bharucha, 1984, who proposed that a non-chordal pitch is heard to be "anchored" to—subordinate to and licensed by—a registrally adjacent chordal pitch. What I propose here is an analogous phenomenon in the rhythmic domain.) It was suggested earlier that this could be captured with a model in which notes are generated in a hierarchical fashion.

Specifically, imagine a position model that generates notes first on beats at levels 3 and 4 (how this is done will be explained below). Notes at level 2 beats are then generated conditional on the "note status" of the neighboring upper-level beats; that is, whether or not they contain notes. There are four possible situations here; there might be: (1) no note on either the preceding or following strong beats (in which case we will call the level 2 beat "unanchored"); (2) a note on the preceding beat but not the following one ("pre-anchored"); (3) a note on the following beat but not the preceding one ("post-anchored"); or (4) a note on both adjacent beats ("both-anchored") (see Figure 11). In the training corpus, we examine every pair of adjacent upper-level (level 3 or 4) beats, classify it as one of the four contexts just described, and then observe whether there was a note on the intervening level 2 beat. This yields the results in Figure 12. We can see, for example, that in a both-anchored context, the probability of a level 2 note is quite high (.713), as it is in a post-anchored context (.889); in a pre-anchored context it is much lower (.265), and in an unanchored context it is lower still (.062). Notice that a pre-anchored note is longer than a post-anchored one; the fact that post-anchored notes are higher in probability than pre-anchored notes thus reflects the avoidance of long notes on weak beats. We repeat this process for level 1 beats, conditional on neighboring stronger beats, and for level 3 beats, conditional on neighboring level 4 beats. (It may seem surprising that the probability of an L3 note in an unanchored context is 1. In fact, there was only one case of an unanchored L3 context—that is, two successive L4 beats with no note on either one—in the entire training set.) The probability of a note on a level 4 beat is simply represented by a single parameter value (.998), and does not depend on neighboring beats. In testing, we compute the probability of a melody by assigning a probability for each B_n, as we would in any position model, but now $P(B_n = \text{note})$ depends on the level of the beat and the note status of the adjacent stronger beats.

Model 5 (Hierarchical Position Model). A decision is made at each beat whether or not to generate a note. The probability of a note at a beat depends on the level of the beat and the note status of the surrounding upper-level beats.

Once again we can use Figure 6 as an example. To assign a probability to this pattern, we first consider the two

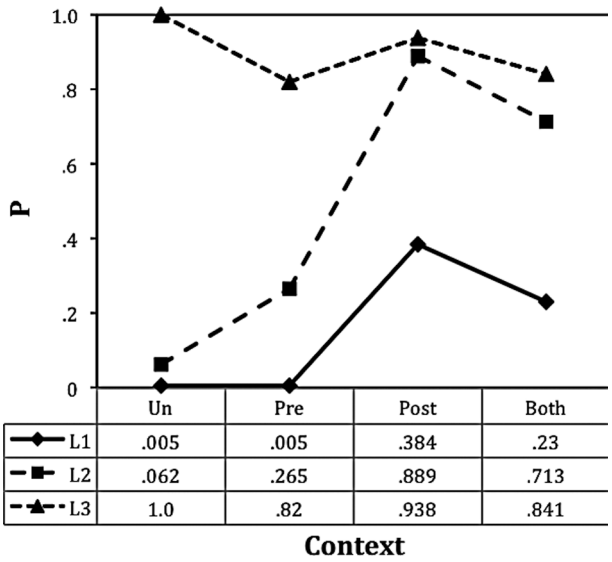


FIGURE 12. Parameters for the Hierarchical Position Model from the Essen training set. "Un," "pre," "post," and "both" are context types (see Figure 11); L1, L2, and L3 are metrical levels. For each combination of context and metrical level, the value shown is the probability of a note-onset. The probability for a note at a level 4 beat is independent of context and is .998.

level 4 beats (position 0 of measure 1 and position 0 of measure 2); both of these beats have notes, yielding probabilities of .998 each. We then consider the L3 beat (position 4); this note is in a "both-anchored" context (since the two neighboring L4 beats both have notes), and position 4 also has a note, yielding a probability of .841. The two L2 beats (positions 2 and 6) are both both-anchored; the second one has a note (.713) and the first does not (1–.713 = .287). Of the four L1 beats (positions 1, 3, 5, and 7), position 1 is pre-anchored (there is a note at position 0 but not at position 2) with no note, yielding 1–.005 = .995; position 3 is post-anchored (there is a note at position 4 but not at position 2) with a note, yielding .384; positions 5 and 7 are both both-anchored with no note, yielding .77 for each. Multiplying these nine probabilities yields the probability of the pattern.

On the Essen test set, Model 5 yields a cross-entropy of 38.76 per song. While this is our best score so far, it is a relatively modest improvement over Model 4, and one might wonder if still further improvement is possible. One possibility is suggested in a model of Raphael (2002). Raphael's model is essentially a duration model, in that it makes decisions for a series of notes. However, in Raphael's model, the probability of a note occurring at a point depends not on the resulting duration but on the note's metrical position; it is also conditioned on the metrical position of the previous note. (We call the note being generated the "consequent" note, and the previous note the "antecedent" note.) For example, the probability of the second note in Figure 6 would depend on its own position within the measure (position 3) and the metrical position of the previous note (position 0). We express this in our final model:

Model 6 (First-Order Metrical Duration Model). A metrical position is chosen for each note, conditional on the metrical position of the previous note.[5]

One could regard this model as a first-order Markov model or "bigram" model of notes, where notes are identified by their metrical positions. (I will sometimes refer to Model 6 simply as the "first-order model," since it is the only first-order model among the ones considered here.) For example, the pattern in Figure 6 features the bigrams 0-3 (the first note is at position 0 in the measure and the second is at position 3), 3-4, 4-6, and 6-0. The probability of a note at a consequent position given an antecedent position, $P(C \mid A)$, is calculated as:

$$P(C \mid A) = count(C, A) \,/\, count(A) \qquad (9)$$

where $count(C, A)$ is the count (number of occurrences) of the bigram and $count(A)$ is the count of notes at the antecedent position. The probability for a melody is the product of $P(C \mid A)$ for all notes in the melody. No probability is assigned for the first note of the melody, but in virtually every song in the Essen corpus, the first note occurs on the first downbeat (recall that partial measures before the first downbeat are excluded); thus this is assigned a probability of 1 (see note 4). Training data for all possible bigrams (antecedent-consequent combinations) are shown in Table 2. (Similar data are presented by Huron, 2006, p. 243. Each consequent position in Table 2 refers to the first possible representative of that position—the one immediately following the antecedent note. That is to say, given an antecedent note at position 0 in m. 1, consequent position 2 in the table refers to position 2 of m. 1 rather than, for example, position 2 of m. 2 or m. 3.[6]) To find the probability of the bigram 0-3, we look up the probability that, given an antecedent note at position 0, the consequent note occurs at position 3; the table shows this probability as .163. The bigrams 3-4, 4-6, and 6-0 yield the probabilities .996, .659, and .686, respectively; multiplying these four values yields the probability of the pattern in Figure 6.

It can be seen that this model captures several of the regularities discussed above. It captures the general preference for notes on strong beats, in that—whatever the position of the antecedent note—there tends to be a preference for the consequent note to be at a relatively strong position. (This can be seen from Table 2: for example, there is generally a higher probability of a consequent note at position 0 than position 1—unless the antecedent note is at position 0.) It also provides a way of capturing the "note length" principle. Given an antecedent note at

[5]Two other theoretical possibilities should be considered briefly. One is a *first-order (non-metrical) duration model*. In such a model, the IOI of each note (its time interval relative to the previous one) would be conditioned on the IOI of the previous note. While this model might achieve a slight improvement over the zeroth-order duration model, it too has no knowledge of meter, and thus seems unlikely to perform very well. Also possible is a *zeroth-order metrical duration model*. Such a model would choose a location for each onset, within (say) a one-measure window after the previous onset, with different probabilities for different positions, but not conditional on the position of the previous onset. (For example, there would be a certain probability of the note occurring at position 2, whether the previous event was at position 1 or at position 7). This model also seems unlikely to perform well. In particular, it has no way of capturing the fact that notes tend to be fairly dense: given a note at any position, the next note is likely to be fairly soon afterwards (this can be seen very clearly in Figure 7).

[6]Consequent positions other than these "immediate" ones are assumed to have a probability of zero. In effect, then, we assign zero probability to interonset intervals of more than one measure; these are extremely rare in the corpus, accounting for less than .03% of all interonset intervals. In testing, we assign such notes the same probability as if they occurred at the same metrical positions with interonset intervals of a measure or less. For this reason the probabilities assigned by the model are slightly higher than they should be.

TABLE 2. Parameters for the First-Order Metrical Duration Model.

	Consequent							
Antecedent	0	1	2	3	4	5	6	7
0	.013	.093	.488	.163	.141	.010	.090	.001
1	.001	0	.999	0	0	0	0	0
2	.003	0	.001	.252	.677	.018	.048	.002
3	0	0	0	0	.996	.002	.001	.001
4	.135	0	.008	0	0	.125	.659	.073
5	.014	0	0	0	0	0	.978	.008
6	.686	0	.002	0	0	0	0	.312
7	1.000	0	0	0	0	0	0	0

position 0, there is a fairly high probability of the consequent note occurring after a relatively long time interval (for example, positions 4 or 6), whereas for an antecedent note at position 1, this probability is much lower (in fact, zero). But both of these principles—the preference for notes on strong beats, and the preference for longer notes on strong beats—are captured by Model 5 as well. Thus it is, *a priori*, not obvious whether this model will be better or worse than Model 5. Testing it on the Essen test set, we find that in fact Model 6 is slightly better; it yields a per-song cross-entropy of 37.36 per song.

No doubt these test results contain some error, due simply to the way the corpus was split into sets for testing and training (though given the large amount of data, it seemed likely that this error would be fairly small). One way to examine this is with K-fold cross-validation. Under this method, the same test is repeated several times, each time using a different portion of the corpus as the test set (and using the remainder of the corpus for training in each case). The original test set contained 20% of the corpus (Test Set 1). The tests reported above were repeated four more times, each time using a different 20% of the corpus as a test set (Test Sets 2 through 5). For each of the six models, the cross-entropy for Test Set 1 was compared to the mean of the cross-entropies for the other four test sets. For each model, the difference was less than 1%; the ranking of the six models was the same as well. This suggests that the results were not greatly affected by the way the corpus was divided into testing and training sets.

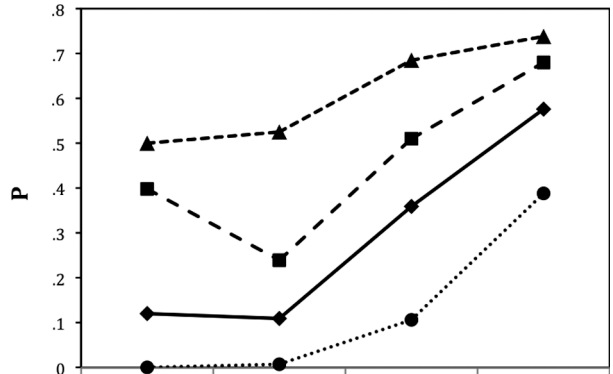
We have now tested six models of rhythm on a corpus of folk melodies. Before drawing any conclusions from this, we apply the same six models to another corpus.

TESTING THE MODELS ON CLASSICAL STRING QUARTETS

The second corpus we will consider consists of all of Haydn's and Mozart's string quartets. Like the Essen collection, these are available in kern notation; we call this the HM corpus. The original corpus contains 309 movements. We again selected only those in 4/4 time. Since the frequency of notes on 16th-note beats is quite a bit higher in the HM corpus, we included movements with notes on 16th-note beats; the models therefore had to be modified to allow this possibility. (We also allowed movements with notes on sub-16th-note beats, though these notes were simply ignored.) This yielded a corpus of 63 movements, or 6,900 measures; 32 movements were put in the training corpus and 31 in the test corpus. Our tests used only the first violin part of each movement. These parts (like many string quartet parts) contain many long stretches of empty measures; these passages are of little interest, and also cause problems for some of the models (especially duration models, since the duration between two onsets may be very large). Thus, all empty measures were deleted from the corpus.

The testing procedure was exactly the same as with the Essen corpus (with the exception that each model was modified to allow notes on 16th-note beats). The results are shown in Table 1. It can be seen that they are broadly similar to the results on the Essen corpus, in terms of the ranking of the different models. One difference is that the Zeroth-Order Duration Model slightly outperforms both the Metrical Position Model and the Fine-Grained Position Model. But the two highest-scoring models are (first) the First-Order Metrical Duration Model, and (second) the Hierarchical Position Model, just as with the Essen data; and the difference in score between these two models is very close to that on the Essen data (the cross-entropy for the first-order model is 4% lower than that of the hierarchical model on the Essen corpus, 5% lower on the HM corpus).

The parameters for the Hierarchical Position Model on the Haydn-Mozart data are shown in Figure 13. There are clear similarities between these values and those drawn from the Essen corpus, shown in Figure 12. In both cases,



	Un	Pre	Post	Both
L0	0	.007	.106	.388
L1	.12	.109	.359	.576
L2	.398	.239	.51	.68
L3	.5	.525	.685	.738

FIGURE 13. Parameters for the Hierarchical Position Model from the Haydn-Mozart training set. The probability for a note at a level 4 beat is .891.

the increase in probabilities as one moves to higher levels reflects the higher probability of notes on stronger beats. Like the Essen corpus, the Haydn-Mozart corpus shows a general preference for both-anchored and post-anchored events over pre-anchored and unanchored ones. There are also significant differences between the two corpora. Post-anchored contexts have higher note probabilities than both-anchored contexts in the Essen corpus (at all levels), while in the Haydn-Mozart corpus, the reverse is true. The higher values for both-anchored contexts in the Haydn-Mozart corpus may be caused partly by long runs of isochronous notes (especially eighth- and sixteenth-notes), which are more common in string quartets than in folk melodies. The Haydn-Mozart corpus also shows a relatively high probability for unanchored notes, especially at the quarter-note level (L2); this indicates a significant presence of syncopation in the Haydn-Mozart corpus, a topic that we will return to below.

COMPLEXITY

It was noted earlier that, assuming all models are equal in prior probability, the one that assigns highest probability to the data is the best model and the one that must be judged most plausible as a model of the compositional process. By this criterion, we must conclude that Model 6—the First-Order Metrical Duration Model—is the best model of the data, with regard to both the folksong corpus and the string quartet corpus. At this point, however, we should reconsider our assumption that all models are equal in prior probability. Cross-entropy indicates the predictive power or "goodness-of-fit" between a model and the data; by all accounts, this is one criterion that should be considered in the model selection process. But other criteria may merit inclusion as well. In particular, one criterion that seems worthy of consideration is complexity. It is generally accepted that, other things being equal, a simpler model is better and more likely to be true (this is just the well-known principle of "Occam's Razor"). To some extent, this factor is addressed by cross-validation: A model that is very specifically tailored to a certain training set is likely to be highly complex, but is also unlikely to perform well on another data set. But studies of model selection (Grünwald, 2004; Pitt, Myung, & Zhang, 2002) have generally argued that some further consideration of complexity is desirable beyond this.

How could the complexity of a model be objectively measured? One standard approach is to define it as a function of the number of free parameters—that is, parameters that are set through training (Pitt et al., 2002). The number of parameters required by each of our six models is shown in Table 1. (For any variable, the number of parameters needed is one less than the possible values of the variable, because all the probabilities must sum to 1. For example, Model 2 has just a single variable with 16 values, hence 15 parameter values. Model 5 has 12 variables—one for each combination of level and context—plus one for level 4; each of the 13 variables has just two values and thus requires just one parameter.) For the moment, let us consider just the Essen corpus. Perhaps not surprisingly, there is a clear trade-off between complexity and predictive power; models with more parameters tend to predict the data better. There is, however, a particularly stark difference between Models 5 and 6 in this regard. While the improvement in predictive power (the reduction in cross-entropy) of Model 6 over Model 5 on the Essen corpus is only about 4%, Model 6 requires more than four times as many parameters as Model 5.

The difference in complexity between Models 5 and 6 becomes even more striking when we consider how the models might be extended. Suppose we modify Models 5 and 6 to allow notes on sixteenth-note beats, as we did with the Haydn-Mozart corpus (see the rightmost column in Table 1). For Model 5, we only need to add four more parameters—one parameter for each of the four contexts for the sixteenth-note level—yielding a total of 17 parameters. For Model 6, however, each measure now has 16 positions, so the number of parameters needed is now $15 \times 16 = 240$. Model 6 now has more than 14 times as many parameters as Model 5. As noted earlier, the difference in predictive power between the two models is not much different between the two corpora; but the addition of the sixteenth-note level greatly increases the difference in complexity between the two models.

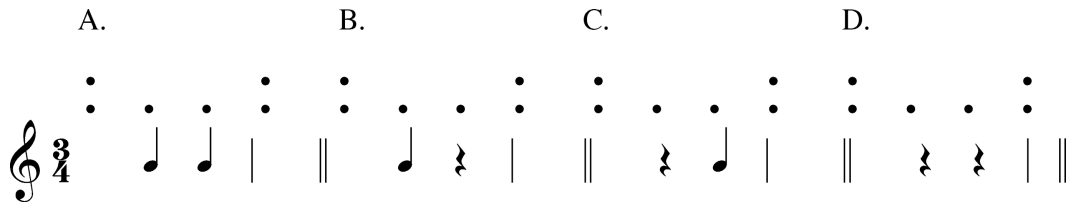


FIGURE 14. Pattern types for the Hierarchical Position Model in 3/4 time.

It is interesting to consider also how the above models might be extended to other time signatures. Consider 3/4 time. In the case of the Hierarchical Position Model, L1 and L0 are essentially unchanged, and the same parameters could be used for these levels as in 4/4 time. (It is not clear whether this would be borne out empirically, but it is at least a reasonable possibility.) With regard to L2, there are now four possible note patterns, rather than just two, that could occur for each of the four contexts (see Figure 14): one could have a note on both L2 beats (A), just the first (B), just the second (C), or neither (D). Thus, 3 parameters would be needed for each context, or 12 altogether for Level 2. Level 3 is now the highest level so only one parameter would be needed for that. A total of $3 + 3 + 12 + 1 = 19$ parameters would thus be needed, six of which would be shared with the 4/4 parameters. In the case of Model 6, by contrast, it is not at all clear how the parameters for 4/4 could be extended to 3/4; it appears that a completely new set of $12 \times 11 = 132$ parameters would be needed. In short, while the Hierarchical Position Model seems to allow some sharing of parameters between time signatures, Model 6 does not, at least not in any obvious way. Here too, then, it appears that Model 5 is a good deal simpler than Model 6.

It seems not unreasonable to argue, then, that when both predictive power and complexity are taken into account, Model 5 is preferable to Model 6, and to the other four models as well. This is really only an opinion, and depends on how the factors of predictive power and complexity are weighed. But it is, at least, one reasonable conclusion.

One might wonder if there was a more objective way of balancing goodness-of-fit and complexity to arrive at a single criterion for model evaluation. One possible solution to this problem lies in the approach known as *Minimum Description Length* (MDL) (Grünwald, 2004; Pitt et al., 2002; Mavromatis, 2005, 2009; Rissanen, 1989). Under this approach, given a body of data, a model can be evaluated by the "description length" that it yields, which includes both the description of the data given the model and the description of the model itself:

$$L_{total} = L(D \mid M) + L(M) \qquad (10)$$

A shorter description means a better model. The term $L(D \mid M)$ can be construed simply as the negative log probability assigned by the model to the data—that is, the cross-entropy. (This relies on a basic principle of information theory: the negative log probability assigned by a model to data is proportional to the number of "bits" required to encode the data under an optimal encoding scheme; a lower probability means more bits, i.e., a longer description.) The $L(M)$ term can then be construed to represent the complexity of the model. (See Mavromatis, 2005, 2009, for an application of MDL to music, though in the domain of pitch rather than rhythm.)

The problem is how the model length $L(M)$ is to be quantified. Several proposals have been put forth for this, but none seems fully appropriate for the current situation. One solution is to define the model description length as proportional or equal to the number of parameters (Papadimitriou et al., 2005). The number of parameters can be added to the cross-entropy to produce an overall measure of "goodness" for each model. The data length favors models fitting the data better, and the model length favors simpler models, as is desired. There is a problem with this approach, however. Recall that cross-entropy is generally expressed in a "per-item" fashion; in Table 1, items are songs, but other units could just as easily be used, such as measures. When comparing cross-entropy values for different models, this makes no difference (as long as the number of items is the same for all models). But in the description length framework proposed above, the item length is crucial: longer items will increase the difference in cross-entropy between models and will thus give data description length (fit to the data) a greater weight than model description length (complexity). Once again, it appears that an arbitrary decision must be made as to how complexity and goodness-of-fit are to be balanced. An alternative approach, used by Mavromatis (2009), is to define data description length as the negative log probability of the entire test set; but then the magnitude of differences in data description length between models (and thus the weight of goodness-of-fit in relation to complexity) will depend on the size of the test set, which also seems wrong.

Another way to define model description length is by considering the average cross-entropy with all possible data sets, the reasoning being that a simpler model is one that achieves a good fit to fewer possible data sets (Grünwald, 2004). This seems counterintuitive, however. In the case of our models, Model 1 achieves roughly an equally good fit to all data sets (all possible rhythmic patterns—since the probability of a note occurring on a beat is roughly .5), which would make it highly complex by the criterion just mentioned. Yet intuitively, it seems that this model is the *simplest* of the ones proposed. (The weakness of Model 1 is not its complexity, but rather, its poor fit to the data.) In short, it seems at present that there is no fully satisfactory method for objectively balancing goodness-of-fit with model complexity.

Discussion

We have now considered six models of melodic rhythm, and have evaluated each of them based on the probability they assign to two musical corpora. On both corpora, the best of the six is the First-Order Metrical Duration Model, which calculates the probability of each note-onset based on its metrical position and the position of the previous note. A close second is the Hierarchical Position Model, which calculates the probability of a note depending on the note status of neighboring strong beats. In terms of complexity, however, the hierarchical model is strongly preferable to the first-order model, as it requires far fewer parameters. In this section we discuss these results and consider some further issues.

First of all, we should consider some questions that may have arisen about the current approach. By the reasoning advanced above, the Hierarchical Position Model is the best model of the data (when both predictive power and complexity are considered) and thus the most plausible model of the composition of common-practice rhythms. But what does this mean? I am *not* suggesting, first of all, that the composition of rhythms was literally "probabilistic," in the sense of involving actual stochastic choices (for example, choices made by flipping a coin or rolling a die). Rather, the use of probabilistic models here—as in other fields—implies that certain aspects of the compositional process are simply unknown and can only be described in an approximate way. Clearly, other factors were involved in the process, and a more complete model would have to specify what these factors were and how they interacted with the hierarchical model. One possibility is that rhythms were constructed by a serial decision process of the kind assumed here, in which decisions were partly guided by the preferences embodied in the hierarchical model (e.g., "prefer post-anchored or both-anchored notes") but also by other considerations—for example, the natural speech rhythm of the text being set (in vocal music), or a preference for larger conventional patterns of various kinds. The parameters of the hierarchical model could then be viewed as measures of the acceptability or "goodness" of various kinds of events. Alternatively, it may be that rhythms were initially generated by a completely different process, such as selection from a set of patterns stored in memory (the "replicative" approach of Cope, 2005, and Gjerdingen, 2007, might play a role here), and that the hierarchical model was then brought to bear as a way of filtering or adjusting the patterns put forth by this initial process (perhaps sometimes rejecting them and sending the process back to the first stage). But under either of these scenarios, the hierarchical model represents a strong empirical claim as to the nature of the compositional process: it implies that patterns were evaluated based on the note status of beats, conditional on the note status of surrounding strong beats. Other models represent alternative claims: for example, the First-Order Metrical Duration Model claims that the goodness of a pattern depends on the metrical position of each note in relation to the position of the previous note. It is important to emphasize, then, that these models do not just represent probabilistic descriptions of the data, but also entail substantive claims about how the data were created.

If the hierarchical model—or any of the other models presented here—really does represent part of composers' musical knowledge, it is natural to ask how this knowledge was acquired. The most plausible answer would seem to be that it was acquired through exposure to music: Composers adjusted the parameters of their internal models (and perhaps even more fundamental aspects of the models themselves) to match the frequency of events in the environment. (This raises the issue of the connection between production and perception, which we return to below.) If this is the case, it might be argued that there is little difference between the view advanced here and the replicative view of Cope (2005) and Gjerdingen (2007). In both cases, composition consists largely of a concatenation of patterns (either directly, or indirectly by some kind of filtering process), whose likelihood of being chosen depends on their frequency of occurrence in the composer's environment. All that differs is the scale—the *granularity*, one might say—of the patterns involved: individual notes or beats (or perhaps, pairs of notes) in my models, larger patterns in Cope and Gjerdingen. While I would accept this view, the issue of granularity is certainly a fundamental one and sets the current models well apart from the replicative theories discussed earlier.

The six models presented above represent a kind of progression: each model adds more fine-grained distinctions, leading to improved cross-entropy performance. In fact, there are really two such progressions, one involving the position models (Models 1, 3, 4, and 5) and the other involving the duration models (Models 2 and 6). No doubt this approach could be extended further. For example, one could posit a position model that conditioned the probability of a note on the note status of adjacent strong beats (like Model 5) but distinguished between different metrical positions of the same strength (like Model 4). Some experiments in this direction are currently underway. Given that five of the six models (all except Model 5) are zeroth-order or first-order Markov models, another natural extension would be to increase the "order" of the models; for example, Model 6 could be refined by conditioning the position of each note on the positions of two previous notes rather than just one. It seems likely that such refinements would result in at least some improvement in performance; however, they also require additional parameters. There will always be a trade-off between complexity and fit to the data; as noted earlier, there seems to be no objective way of optimizing this trade-off.

This leads to a further question: How good are these models? Our probabilistic method allows us to compare the models to one another in terms of predictive power, but it gives no way of assessing them in an absolute sense. How good could a probabilistic model of the Essen test set possibly be? One way to approach this issue is from the point of view of information theory. Our simplest model, the Uniform Position Model, assigns a per-song cross-entropy (negative log probability) to the corpus of 62.37. This could be regarded as the amount of uncertainty in the data from the point of view of a model that knows nothing other than the proportion of beats that have note-onsets; it could thus be seen as a kind of baseline. By contrast, the Hierarchical Position Model assigns a cross-entropy of 38.76; the difference between the hierarchical model's score and the baseline score indicates the amount that the uncertainty of the data has been reduced by the hierarchical model, and the model's score indicates the amount of uncertainty that remains. The question is, how much further reduction a model could reasonably be expected to obtain.

No doubt, a large amount of the uncertainty in the data is due simply to individual variation between melodies. Clearly, there were factors in the composition of melodies that varied from one melody to another—depending on the composer, the composer's state of mind and goals, the text being set, and so on. (Otherwise, all the melodies in the corpus would be the same.) This individual variation means that it is theoretically impossible for a model to predict the rhythm of a melody with probability of 1, because all melodies are different. The models being considered here do not even attempt to account for this individual variation. Thus, there is some kind of limit on the reduction in uncertainty that a model can be expected to achieve.[7]

As well as uncertainty in the data due to individual variation, however, there may also be systematic regularities, beyond those captured by the models presented above, that a rhythmic model might be expected to account for. There may for example be conventional patterns, perhaps of a measure or more in length, that occur often (again, the replicative approach comes to mind). There are also principles of large-scale rhythmic organization that the current models do not capture—for example, the preference for four- or eight-measure phrases, which certainly has implications for rhythm. A further regularity is the preference for rhythmic repetition within songs: once a rhythmic pattern occurs, it is likely to occur again. No doubt, a model that incorporated these factors into its predictions could achieve a substantial improvement over the models presented here. Just how much uncertainty could be reduced in this way, and how much is due to individual variation between melodies, remains an open question.

Modeling Rhythm Perception

Our assumption has been that the composition of common-practice rhythms involves general principles of some kind; the models we have considered are hypotheses as to what those principles might be. It seems clear that the *perception* of common-practice rhythm, too, is governed by general principles. In this section we consider some ways that the rhythmic models proposed above might be evaluated with regard to perception. The perceptions I will examine are those of present-day Western listeners (as represented by recent experimental work and by my own intuitions). It is not obvious that the

[7] As mentioned earlier, the theoretically optimal model would be a "cheating" model that assigned a probability of $1/N$ to each of the N melodies in the test set. As just mentioned, even this model does not achieve a probability of 1, due to variation between songs. But no legitimate model could be expected to obtain even this result. The melodies in the test set are drawn from a much larger population of possible melodies—other existing folk melodies that simply were not in the corpus, as well as the innumerable possible melodies that are within the style and could have been written but were not. Any honest model of European folk melodies would need to leave some probability mass for these other actual and possible melodies.

principles governing the perception of rhythm among this population would be the same as those governing the creation of common-practice rhythms—especially since the music on which our models were tested was mostly written in a different historical context (pre-twentieth-century Europe). But in fact, as we will see, the models that perform best with regard to compositional practice also receive the strongest empirical support as models of perception.

REPRESENTING METER

Much recent experimental work on rhythm perception has been concerned in some way with meter. This work has shown that meter plays a role in rhythm perception in a variety of ways. Patterns that strongly support a regular meter are encoded more easily and perceived as less complex than those that do not (Povel & Essens, 1985); the same melody presented in two different metrical contexts can seem quite different (Povel & Essens, 1985; Sloboda, 1985). Meter also affects expectation; for example, the pitch of a note can be more accurately judged when it occurs at a metrically expected position (Jones, Moynihan, MacKenzie, & Puente, 2002). Meter affects performance as well; notes in metrically similar positions are more likely to be confused in performance errors (Palmer & Pfordresher, 2003), and aspects of expressive performance—timing and dynamics—also betray a subtle but important effect of meter (Drake & Palmer, 1993; Sloboda, 1983). Elsewhere I have argued that meter plays an important role in the perception of harmony and phrase structure, and also affects the perception of repeated patterns in music: when considering two segments as possibly similar or "parallel," we are strongly biased towards pairs of segments that are similarly placed with respect to the meter (Temperley, 1995, 2001).

In light of all the evidence for the central role of meter in rhythm perception, a perceptual model of rhythm must clearly represent meter in some way in order to be plausible. Models that make no distinction at all between different metrical positions—such as Models 1 and 2 above—seem fatally flawed in this regard. Models 3 and 5, by contrast, clearly distinguish between levels of metrical strength. As for Models 4 and 6, these models are clearly superior to Models 1 and 2, in that they distinguish between different metrical positions and recognize the similarity between beats of the same metrical position. However, they are in a sense *too* fine-grained: they do not capture the similarity between different positions of the same metrical strength, such as positions 1, 3, 5, and 7 in Figure 6. Of course, some indication of metrical strength could be added to these models. But it is surely an advantage of Models 3 and 5 that they already represent metrical levels explicitly, without the need for any further metrical information.

SYNCOPATION

Another possible way of evaluating rhythmic models with regard to perception concerns their handling of syncopation. Syncopation is a familiar and widely used concept in discourse about rhythm, but is difficult to define precisely. The New Harvard Dictionary of Music (Randel, 1986) defines syncopation as "a momentary contradiction of the prevailing meter or pulse." This definition seems fairly close to the usual usage of the word. For example, a rhythmic pattern such as that in Figure 1B would normally be described as highly syncopated, because it seems to go against the meter in which it is notated. However, this definition lacks rigor: what exactly does it mean for something to be a "contradiction" of the meter? Huron and Ollen (2006) define a syncopation as "the absence of a note onset in a relatively strong metric position compared with the preceding note onset" (p. 212). This definition is more rigorous, but seems imperfect. In Figure 9C, the second note is a weak-beat note with no onset on the following strong beat, but I think few would consider this pattern to be syncopated; at most, it is a syncopation of a very mild sort.

I propose an alternative definition of syncopation that I would argue is both rigorous and intuitively satisfactory: a syncopated rhythm is one that is low in probability given the prevailing meter (by the norms of common-practice rhythm). This is similar to the Harvard Dictionary of Music definition: If a rhythm is low in probability given a meter, it contradicts that meter, in that the meter is likely to be low in probability given the rhythm. (In Bayesian terms, if $P(\text{note pattern} \mid \text{meter})$ is low, then $P(\text{meter} \mid \text{note pattern})$ is also likely to be low. We return to this Bayesian view of rhythm perception below.) One could hardly dispute that Figure 1B is much less likely in the context of common-practice rhythmic norms than Figure 1A. We have not yet said exactly how $P(\text{note pattern} \mid \text{meter})$ will be determined; we will return to this issue below. We should note right away, however, that it will matter greatly what kind of music is used to set the model's probabilities. Much twentieth-century popular music is highly syncopated; in such music, we would expect the probabilities of some syncopated patterns to be quite high. What makes a pattern seem syncopated, I would argue, is that it is low in probability in relation to the norms of common-practice rhythm.

By the construal of the term proposed above, syncopation is closely related to rhythmic complexity. It seems plausible to suggest that the complexity of a rhythmic pattern is related to its probability; less probable rhythms seem more complex. (Figure 1B above seems more complex

than Figure 1A, for example.) However, complexity also may be affected by factors other than syncopation, notably the amount of repetition in a pattern. A repetitive pattern is less complex—and one could have a syncopated pattern that was extremely repetitive. So complexity and syncopation are related, but not equivalent. It might be said, perhaps, that syncopation represents the aspect of rhythmic complexity that does not relate to repetitiveness.

In order to flesh out our probabilistic definition of syncopation, we need a way of calculating *P*(note pattern | meter). In fact, we already have addressed this problem. Each of the models presented earlier can be used to calculate *P*(note pattern | meter); this is precisely the quantity that was used to evaluate the models with regard to common-practice rhythmic composition. Therefore, each model yields a measure of syncopation that could be applied to any given rhythmic pattern. The question is, which of these measures of syncopation corresponds most closely with our intuitive understanding of the term?

We can dispense quite easily with Models 1 and 2, as they do not consider meter at all. (By these models, a rhythmic pattern is equally likely given any meter, and thus cannot be said to contradict one meter more than any other.) With Models 3 and 4, we gain an awareness of distinctions between different positions within the measure, and this allows some recognition of syncopation. However, the ability of these models to recognize syncopation is limited. It can be seen that neither model pays any attention to the context of a note, only to its metrical position; and in some cases context is important. Return once more to Figure 5. According to Models 3 and 4, Figures 5A and 5B are equivalent: Each one has two notes at position 0 in the measure, and one note each at positions 3, 4, and 6. Yet Figure 5B is surely more syncopated than Figure 5A. Figure 5B features a note on a weak (eighth-note) beat with no note on either adjacent beat, which (in terms of the Harvard Dictionary definition) seems to contradict the meter; in Figure 5A, the weak-beat note is followed by a note on the strong beat.

We now consider Models 5 and 6. Both of these models are of course sensitive to metrical position. And both models also are sensitive to the context of a note, though in different ways. The cross-entropy assigned by each model to Figures 5A and 5B (assuming the Essen parameters for both models) is shown in Table 3. (We assume that there is a note on the downbeat of the following measure; this is necessary, since Model 5 requires a span with downbeats at both ends.) Both models succeed in assigning a lower cross-entropy (higher probability) to Figure 5A than to Figure 5B, but for different reasons. For Model 6, Figure 5B is low in probability because of the transition between the second note and the third; given a note at position 3 of the measure, it is extremely unlikely that the following note will be at position 6 (the probability of this is just .001). For Model 5, Figure 5B is low in probability because the second note is an "unanchored" Level 1 note—there is no note on either the previous or following strong beat (this has a probability of .005).

TABLE 3. Cross-Entropy Assigned to Figures 5A and 5B by the Hierarchical Position Model and the First-Order Metrical Duration Model.

	Hierarchical Position Model	First-Order Metrical Duration Model
Figure 5A	8.13	6.95
Figure 5B	12.00	13.06

Can we distinguish between Models 5 and 6 in their handling of syncopation? The patterns in Figure 15 offer a possible way. For Model 6, what matters is the metrical "bigrams"—positions of pairs of adjacent notes. Figure 15A has the bigrams 0-4, 4-5, 5-0, 0-5, 5-6, 6-0; Figure 15B has 0-5, 5-0, 0-4, 4-5, 5-6, 6-0. It can be seen that the two patterns contain exactly the same bigrams; their order is different, but this is irrelevant for Model 6. Thus, for Model 6, these two patterns are equivalent in probability. For Model 5, however, they are not; there are several differences between the two patterns, such as the fact that the second contains an unanchored L1 note (the second note), while the first does not. Because of these differences, Model 5 assigns different probabilities to the two patterns, assigning a lower cross-entropy to Figure 15A (12.44) than Figure 15B (14.38). It seems to me this is intuitively correct; Figure 15B *does* seem more syncopated than Figure 15A. While this may be a rather contrived example, it suggests that are at least some situations where Model 5 provides a better model of syncopation than Model 6.[8]

A small test was done to assess the models' ability to predict perceived rhythmic complexity. Povel and Essens' classic study (1985) presents an experiment in which subjects heard 35 short rhythmic patterns (shown in their Table 2) and had to reproduce them. The patterns were constructed from permutations of the durations 1/1/1/1/1/2/2/3/4, with the "4" at the end; as the authors

[8]Although Model 5 judges Figure 15B as more syncopated than Figure 15A, it does not seem to do so for the right reason. What seems syncopated about Figure 15B is the unanchored L1 note (the second note in the pattern); but in fact, using the Essen parameters (as we are here), Model 5 gives the same probability for an unanchored L1 note as a pre-anchored one (.005).

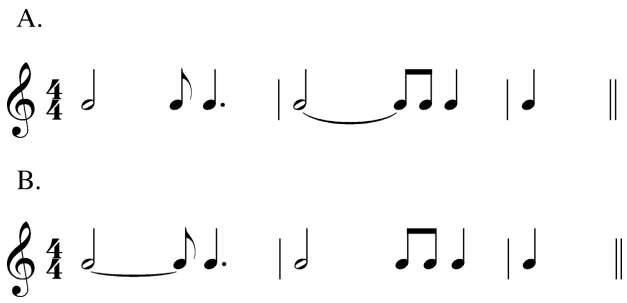


FIGURE 15. These two patterns are equivalent under the First-Order Metrical Duration Model, but not under the Hierarchical Position Model.

suggest, this long interval at the end of the pattern tends to induce a meter with beats at the beginning and end of the interval (though they caution that some of the more complex patterns may not have clearly induced any meter). Two of their patterns are shown here in Figure 16. With patterns as short and constrained as these, there clearly is not much room for repetition; thus, by the current view, differences in complexity should be largely due to syncopation. Povel and Essens examined the average deviation (in relation to the correct interonset intervals) in subjects' reproductions of each pattern and present this in a graph (their Figure 9); this can be taken as a measure of reproduction accuracy, or more precisely *in*accuracy, and hence, as an indication of complexity. As the original data are no longer available (Povel, personal communication), I read the approximate values (rounding off to the nearest 6 ms) from the graph itself. I then encoded the patterns (treating "1" as the eighth-note and assuming the meter indicated in Figure 16), submitted them to the six models to obtain the cross-entropy for each pattern, and looked at the correlation between each model's cross-entropy judgments and reproduction inaccuracy. The results are shown in Table 4. For Models 1 and 2, all 35 patterns receive the same cross-entropy, thus the correlation with the experimental data is zero. The remaining four models exhibit the same ordering found in modeling the Essen corpus, except that the hierarchical model outperforms the first-order model. While further testing is necessary, this small experiment gives some encouraging support for the hierarchical model with regard to the perception of rhythmic complexity.

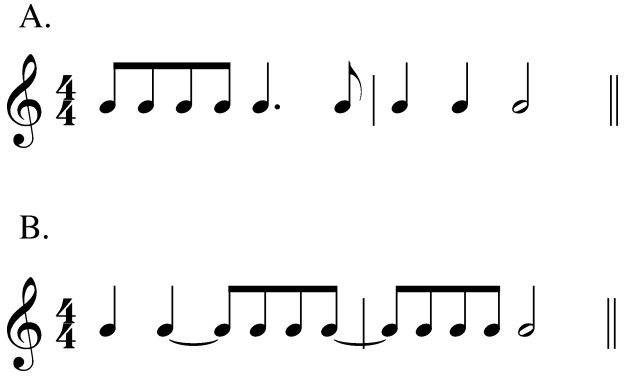


FIGURE 16. Two rhythmic patterns from Povel and Essens (1985), Table 2: (A) no. 1 and (B) no. 35.

TABLE 4. Correlations between Rhythmic Models' Cross-Entropy Judgments and Reproduction Inaccuracy for 35 Patterns Used in Povel & Essens (1985).

Model	*r*
1. Uniform Position Model	0
2. Zeroth-Order Duration Model	0
3. Metrical Position Model	.61
4. Fine-Grained Position Model	.65
5. Hierarchical Model	.76
6. First-Order Metrical Duration Model	.71

MODELING METER-FINDING

One very basic and important aspect of rhythm perception is the identification of meter. This process has been the subject of a huge amount of research, both experimental and computational (for surveys, see Temperley, 2001, and Gouyon & Dixon, 2005). In recent years, several authors have approached this problem from a probabilistic viewpoint (Cemgil, Desain, & Kappen, 2000; Raphael, 2002; Temperley, 2007). In probabilistic terms, the meter-finding problem can be defined as the problem of determining the most probable metrical structure given a pattern of notes. Bayes' Rule then tells us that

$$(\text{meter} \mid \text{note pattern}) \propto P(\text{note pattern} \mid \text{meter}) \times P(\text{meter}) \qquad (11)$$

The meter that maximizes the right side of this expression will be the most probable meter given the note pattern (Temperley, 2007).

As noted with regard to syncopation, the first term on the right side of expression (11) is the quantity used in the experiments presented earlier to test models of compositional practice: each of our models offers a way of calculating the probability of a note pattern given a meter (though the parameters of the models were only specified for 4/4 meter). As for the second term, this captures the fact that some meters may simply be more probable than others: For example, if 5/4 meter occurs very rarely, we should probably not infer it unless there is overwhelming evidence in its favor.

If we assume that listeners can generally infer the correct meter for common-practice melodies, we could test

our six models by converting them into meter-finding models and testing their ability to identify meters correctly. (Models 1 and 2, having no knowledge of meter, would clearly fail at this task.) This would certainly be of interest, and would provide another way of evaluating the relative plausibility of the various models; it would, however, be a major undertaking. A meter-finding model must determine not only the time signature of the meter, but also the phase: that is, if a melody is in 4/4 time, does the first note occur on the first eighth-note beat of the measure, the second, or some other beat? For each model, the probabilities of these different time signatures and phases would have to be determined. "Likelihood" functions—defining $P(\text{note pattern} \mid \text{meter})$—would also have to be defined, not only for 4/4 time (as we have already done) but for other time signatures as well. In addition, for a meter-finding model to be at all realistic, it should not assume that the input is "quantized"—played with perfectly regular timing; it should allow for beats to be somewhat irregularly spaced, as they are in real music.

In fact, several prior meter-finding models could be seen as implementations of the generative models proposed above (or the principles behind them). The basic principle of Model 3—that the probability of a note at a beat depends on its metrical strength—is implicit in many meter-finding models, and is reflected explicitly in the probabilistic models of Cemgil et al. (2000) and Temperley (2007). In recent work, I have proposed a meter-finding system based on the Hierarchical Position Model (Temperley, 2009). The First-Order Metrical Duration Model has also been implemented in a meter-finding system by Raphael (2002). However, the implementation and testing of these models is so different that it is not really possible to compare them. Suffice it to say that the compositional models presented earlier (at least, Models 3, 4, 5 and 6) could be used for meter-finding, and this might offer another way of deciding between them as models of perception.

The Bayesian view suggests a profound connection between rhythm production and perception. By this view, in order to infer the metrical structure for a rhythmic pattern, the listener must know the probabilities of different metrical structures and also the probabilities of different patterns given those structures. Of interest in this connection is a study by Sadakata, Desain, & Honing (2006), which uses a Bayesian perspective to model the perception of rhythm. The study focuses on three-onset patterns, in which the first and third onsets fall on tactus beats (assuming tactus intervals of one second); the placement of the second onset divides the interval into two notes. We can represent such patterns by identifying the length of the first note as a proportion of the larger interval; thus 1/4 implies the rhythm sixteenth-note / dotted-eighth. Corpus analysis was used to find the prior probability of different beat divisions in notated music, and production data were analyzed to find the distribution of performed rhythms given different notated rhythms. Using Bayesian logic, these data were then used to predict rhythmic perception, and the model's predictions were compared to data from perception experiments in which subjects were played three-onset patterns and asked what notation they implied; the predictions fit the data very closely. The model explains certain apparent asymmetries between production and perception data. For example, the pattern 2/5 is typically performed at close to its "correct" timing, but a performed rhythm with this timing is much more likely to be perceived as 1/3; according to the model of Sadakata et al., this is because a 1/3 tactus division is much higher in prior probability than 2/5.

It could be said that the model of Sadakata et al. (2006) entails a probabilistic model of rhythm composition, in that different probabilities are assigned to different notated tactus divisions. The model is somewhat similar to our hierarchical model, in that it assumes a context of two strong beats with note onsets and states the probability for certain patterns occurring on low-level beats in between. The fit between the model and perception data is certainly impressive and is a strong general validation of the Bayesian approach to rhythm. The model is clearly limited in that it only allows one onset between two strong beats; this limitation could be addressed by adding further patterns, but eventually this could lead to a huge proliferation of patterns (and parameters). As the authors themselves note, "Surely listeners do not memorize a huge number of distributions for different complex rhythms. Perception of complex rhythm must be based on simple rhythm in a principled way." The models proposed above offer some possibilities as to how this might be done.

Conclusions

The main aim of this study has been to evaluate models of common-practice rhythm, focusing especially on the compositional processes involved in their creation. Six models were considered, and each one was evaluated as to the probability it assigned to the rhythms in two corpora of common-practice pieces in 4/4 time: a corpus of European folk songs and a corpus of Mozart and Haydn string quartets. Two models clearly performed the best at this task: the Hierarchical Position Model, which generates notes in a hierarchical fashion conditional on the note status of neighboring strong beats, and the First-Order Metrical Duration Model, which chooses metrical

positions for notes based on the current position and the position of the previous note. The first-order model performed slightly better than the hierarchical model with regard to the sheer probability assigned to the corpora, but is also significantly more complex (requiring more parameters), particularly when a 16th-note level is added and when the models are extended to other time signatures.

We also considered these models as models of perception. Given the crucial importance of meter in many aspects of rhythm perception, models that explicitly encode metrical levels (such as the hierarchical model) seem to have an advantage over those that do not (such as the first-order model). With regard to syncopation, it was argued that both the hierarchical model and the first-order model predict syncopation quite well but that the hierarchical model may have a slight edge in this regard; a small test on experimental data showed an advantage for the hierarchical model. Finally, I noted that probabilistic models of rhythm may be construed as meter-finding models, and that this provides another way of evaluating them. While both the hierarchical model and the first-order model have been incorporated into meter-finding models and seem quite successful, there has not yet been any attempt to compare them; this would be an interesting project for the future.

My conclusion is that, on balance, the hierarchical model is the most plausible of the models we have considered, with regard to both composition and perception. This is—as already stated—really a matter of opinion, and depends on how the various sources of evidence considered above are weighted. Clearly, both the Hierarchical Position Model and the First-Order Metrical Duration Model are serious contenders with many virtues; further work is needed to decide between them conclusively.

As well as the evaluation of specific models, a second, broader aim of this study has been to argue that the development and evaluation of models of composition is an appropriate and worthwhile project for the field of music cognition. Models of composition can be evaluated, I have suggested, by examining how well they predict compositional data, and probabilistic methods are extremely well-suited to this task. As explained earlier, no claim is being made that any of the models proposed here are *complete* models of the compositional process, nor do they imply any assumption that "composition is probabilistic" (whatever that might mean). But they nonetheless entail strong and substantive claims about the compositional process; and probabilistic methods allow for the testing of these claims in quantitative, objective ways.

Author Note

I am grateful to Stephen McAdams, Panos Mavromatis, Peter Pfordresher, Geraint Wiggins, and an anonymous reviewer for helpful comments on earlier drafts of this article.

Corresponding 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

Bharucha, J. J. (1984). Anchoring effects in music: The resolution of dissonance. *Cognitive Psychology*, *16*, 485–518.

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

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

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

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

Drake, C., & Palmer, C. (1993). Accent structures in music performance. *Music Perception*, *10*, 343–378.

Gjerdingen, R. O. (2007). *Music in the galant style*. New York: Oxford University Press.

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

Grünwald, P. (2004, June 4). *A tutorial introduction to the minimum description length principle* [Online tutorial]. Retrieved from http://arxiv.org/abs/math.ST/0406077

Huron, D. (1990). Increment/decrement asymmetries in polyphonic sonorities. *Music Perception, 7*, 385–393.

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

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

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

Huron, D., & Ollen, J. (2006). An empirical study of syncopation in American popular music, 1890–1939. *Music Theory Spectrum*, *28*, 211–231.

Jones, M. R., Moynihan, H., MacKenzie, N., & Puente, J. (2002). Temporal aspects of stimulus-driven

attending in dynamic arrays. *Psychological Science, 13*, 313–319.

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

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

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

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

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

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

Papadimitriou, S., Gionis, A., Tsaparas, P., Väisänen, R., Mannila, H., & Faloutsos, C. (2005). Parameter-free spatial data mining using MDL. In H. Kargupta, J. Srivastava, C. Kamath, & A. Goodman (Eds.), *Proceedings of the 5th International Conference on Data Mining* (pp. 346–353). Philadelphia, PA: Siam Books.

Pearce, M. T., & Wiggins, G. A. (2004). Improved methods for statistical modelling of monophonic music. *Journal of New Music Research*, *33*, 367–85.

Pitt, M., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.

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

Randel, J. (1986). *The new Harvard dictionary of music.* Cambridge, MA: Harvard University Press.

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

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

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

Schaffrath, H. (1995). *The Essen folksong collection* [Online database]. D. Huron (Ed.). Retrieved from http://essen.themefinder.org/

Sloboda, J. A. (1983). The communication of musical metre in piano performance. *Quarterly Journal of Experimental Psychology*, *35*, 377–396.

Sloboda, J. A. (1985). *The musical mind*. Oxford: Clarendon Press.

Temperley, D. (1995). Motivic perception and modularity. *Music Perception*, *13*, 141–169.

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

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

Temperley, D. 2009. A unified probabilistic model of polyphonic music analysis. *Journal of New Music Research*, *38*, 3–18.