The Predictive Benefit of Order Increases in Melodic N-Gram Models

David Temperley[1(✉)] and Matt Chiu[2]

[1] Eastman School of Music, University of Rochester, Rochester, NY 14618, USA
dtemperley@esm.rochester.edu
[2] Baldwin-Wallace University, Berea, OH 44017, USA

Abstract. An important question about n-gram models is, how much gain in predictive power is achieved as n-gram order is increased? We offer a new method for approaching this question—one that avoids the difficulties that arise from testing on unseen data, and also avoids the overestimation of predictive benefit that can occur with small data sets. For each *n*th-order context, we choose a sample of tokens of the "parent" (*n*-1)-order context whose count exactly matches that of the *n*th-order context, and compare the conditional entropy of the two contexts. Using corpora of folk songs, classical themes, and hymn tunes, we provide the first unbiased estimate of the predictive benefit of 1st- through fourth-order melodic n-gram models. Comparing predictive benefits for different intervals within a single n-gram order, we find patterns that reflect the influence of basic principles of melodic structure: pitch proximity, step inertia, and range constraints.

Keywords: melody · entropy · n-gram models · probabilistic models

1 N-Gram Models

N-gram modeling is a method for defining the probabilities of events in a sequence. In an n-gram model (also known as a Markov model), the probability of an event is conditioned on the previous few events. The number of previous events considered is the order of the model: a 0th-order model considers no context; a first-order model considers the previous event; a second-order model, the two previous events; and so on. N-gram models have a long history in both language and music research (Shannon, 1951; Youngblood, 1958). In music, the most notable application of n-gram models in recent years is Pearce's IDyOM system (2018), building on the previous work of Conklin and Witten (1995). The IDyOM system has been used to model human expectation judgments (Pearce & Wiggins, 2006), the perception of segmentation boundaries (Pearce et al., 2010), and judgments of complexity (Sauvé & Pearce, 2019), among many other applications. N-gram models can also be used to model sequential data itself: the probability that an n-gram model assigns to a body of data can be taken to indicate the "fit" of the model to the data. Varying the features and parameters of an n-gram model to optimize its fit can provide insight into the constraints that influenced the compositional process (Schubert & Cumming, 2015).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
M. Montiel et al. (Eds.): MCM 2026, LNAI 16609, pp. 487–501, 2026.
https://doi.org/10.1007/978-3-032-27827-2_29

In general, higher-order n-gram models perform better than lower-order ones—as long as sufficient data is available to train them (we will return to this point). The more context the model has to look at, the more it is able to intelligently guess the next event. We could say, then, that increasing the order of an n-gram model generally yields a *predictive benefit.* In modeling words in language, increases up to seventh-order have been shown to yield improvement (Goyal et al., 2009); such models can make use of semantic and syntactic information provided by fairly distant words. In music, the situation is not so clear. It seems likely that the probability of a note is affected by the previous one or two notes, but the effect of more distant context is not so obvious. While we would expect there to be *some* predictive benefit of (say) a third-order model over a second-order one, the magnitude of this benefit has, to our knowledge, never been shown. In this paper we present a novel method for determining the predictive benefit of increases in n-gram order on musical (melodic) data. Our tests also yielded a further, unexpected, result: The predictive benefit of an *n*th-order model over an (*n*-1)-order one varies significantly, depending on the interval at the beginning of the *n*th-order context.

2 Methodology

An n-gram model is typically tested by defining the conditional probabilities of events in a training set and then using those probabilities to define the probability of a separate test set given the model. We use essentially that approach here, except that we use the same dataset for training and testing—an unconventional approach that will be defended below. The probability of a dataset of melodies given an n-gram model can be defined as:

$$P(\text{dataset}|\text{ model}) = \prod_n P(E_n|\text{context}) \tag{1}$$

where E_n is the *n*th event (i.e. note) in the dataset. We take the logarithm (to avoid the very small numbers that arise from multiplying many probabilities together), divide by the number of events to produce a "per-event" value, and make it negative so that lower probabilities yield higher values. This produces a measure of the uncertainty or unexpectedness of the dataset, known as *conditional entropy* (CE):

$$\text{CE} = -1/N \sum_n \log_2 P(E_n|\text{context}) \tag{2}$$

where N is the total number of events. The negative log probability of a single event is known as its *surprisal*; CE is the average surprisal of all events. To understand how $P(E_n \mid \text{context})$ is calculated, we need to think of identical events as being grouped into categories (or "target types") T; identical contexts are grouped into context types C. $P(E_n \mid \text{context})$ is the proportion of occurrences of that particular context type that are followed by the target type of E_n, or $P(T \mid C)$. We can then express CE as

$$\begin{aligned} \text{CE} &= -\sum_C P(C) \underbrace{\sum_T P(T \mid C) \log_2 P(T \mid C)}_{\text{CE contribution of context } C} \\ &= -\sum_{C,T} P(C, T) \log_2 P(T \mid C) \end{aligned} \tag{3}$$

(This is a more standard definition of CE than (2), though the two are equivalent.) In the current case, the context will be some number of events immediately preceding the target event, the number depending on the n-gram order being used. Every context type makes a contribution to the CE (its "CE contribution"), which can be viewed as the CE (or average surprisal) of targets given that context type; the overall CE is a weighted mean of these entropy contributions, each one weighted by the frequency of its context. (This is represented by the first line of Eq. 3.) If the events following a context type are distributed across many target types, all of them will be low in probability, and the CE contribution will be high; if one particular target type is much more common than the others in that context, the CE contribution will be lower. In the extreme case, if a particular C is always followed by a particular T, $P(T \mid C) = 1$, and C's entropy contribution will be $-\log_2(1) = 0$, meaning that the target is completely predictable given the context.

In the case of the 0^{th}-order model, there is no context, and conditional entropy simply becomes *entropy*:

$$\text{entropy} = -\sum_{T} P(T)\log_2 P(T) \tag{4}$$

One could also view this as a special case of conditional entropy in which all events have the same context; we find it convenient to do that here.

CE is an indicator of the probability of the dataset given the model (higher CE indicates lower probability). Suppose we measure the CE of a corpus of melodies using both a first-order n-gram model and a second-order model. If the CE is the same or nearly the same in both cases, this will suggest that expanding the context from one to two preceding events adds little or no predictive power. If, however, the second-order model yields lower CE than the first-order model, that will point to a predictive benefit for the second-order model over the first-order one.

In most studies using n-gram models, one sample of the data is used to train the model (defining the $P(T \mid C)$ terms in the second line of Eq. 3 above) and a separate sample is used to test it (defining the $P(C,T)$ terms). This is partly because n-gram models—when intended for practical purposes such as speech recognition, or as realistic models of human processing—usually need to be usable on data that has not been seen in training, so it is important to examine their performance in such situations. In the current case, however, the model is designed to answer a basic research question; it is not necessary that it be usable on unseen data. Thus, for each order-specific model, a single dataset is used both for setting the model's parameters and for measuring CE. Using the same sample for training and testing simplifies the process considerably. If separate training and testing samples are used, there is a danger, especially with higher orders, that the model will encounter contexts that were not seen in training and whose probabilities are therefore undefined, or target events following a certain context that were not seen following that context in training, yielding a probability of zero and a CE of infinity. These "sparse data" problems with higher-order models are usually solved by combining them with lower-order models, a process known as "smoothing." Notably, Pearce and Wiggins (2004), who examine the predictive power of different n-gram models on melodic data, use separate training and testing sets and also use smoothing.

In the current case, smoothing is not only unnecessary but undesirable, since the aim is precisely to compare CE between different n-gram orders.

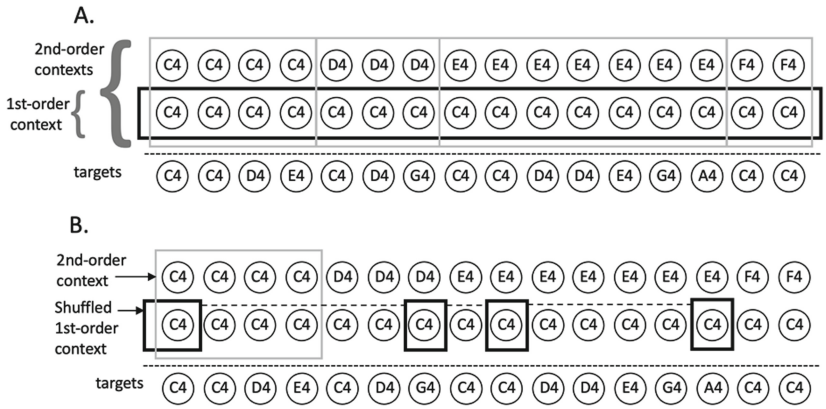

Fig. 1. An illustration of the shuffled-context procedure, using hypothetical data. (A) Normally, the entropy of a first-order context is calculated as the average surprisal of targets given the context (the black rectangle shows the tokens that would be considered in such a calculation); the same is done for second-order contexts (the grey rectangles). (B) Under the shuffled-context procedure, for each second-order context, a "shuffled" first-order (parent) context is created by randomly selecting from the parent context the same number of tokens as in the second-order context.

One problem with training and testing on the same dataset is that it is likely to underestimate the true conditional entropy of the population from which the dataset was extracted. Consider a context that occurs only once in the dataset. Whatever event follows that context will be assigned a probability of 1, and the contribution of that context to the total CE will be zero—though given a larger dataset from the same population, the CE of that context might be much greater than zero. If a context occurs twice, its maximum possible CE is 1; if it occurs five times, it is 2.3. We will refer to this as the "small count problem" (it is also sometimes called "overfitting"). This problem is especially likely to arise with higher n-gram orders; since the number of possible contexts increases as n-gram order increases, the counts of those contexts will decrease, possibly yielding many contexts with very small counts. Thus, the small count problem is likely to cause the predictive benefit of higher-order models to be overestimated. The severity of this problem will depend on the amount of data available; our initial experiments with the datasets considered below suggested that it was quite severe.

Here we propose a novel solution to the small count problem (see Fig. 1). Suppose we wish to compare the CE of an nth-order context to that of its (n–1)-order parent. Rather than using all the available tokens of the parent context, we take a random sample whose count exactly equals that of the child context. In this way, we eliminate the small count effect. We call this the "shuffled-context" procedure; one can think of it as shuffling the parent context tokens, randomly assigning each one to a child context for comparison. We can then calculate the CE of the parent context as a weighted mean of the context CE's produced by this shuffling procedure. The overall CE of the shuffled parent model

is a weighted average of its context CE's, as it would normally be; this can be compared to the CE of the child model. This sampling procedure can be repeated multiple times with the parent contexts to reduce random noise. (This cannot be done with the child contexts, since we are already using all of the available data.)

Table 1. CE of n-gram models on the text corpus (letters)

Model	CE	Shuffled CE of (n-1) order	Predictive benefit (% improvement of nth order over shuffled (n–1) order)
0^{th}-order	4.142	—	—
1^{st}-order	3.443	4.140	16.8
2^{nd}-order	2.806	3.416	17.9
3^{rd}-order	2.113	2.688	21.4
4^{th}-order	1.560	1.892	17.6

As a simple illustration of our method, let us consider letters in written English. We took 2000 sentences (about 245,000 letters) from a widely used corpus of text from the *Wall Street Journal* (Marcus et al., 1993); we removed all punctuation and numbers, but not spaces, and converted upper-case letters to lower-case, leaving an alphabet of 27 symbols. We created 0^{th}-, 1^{st}-, 2^{nd}-, 3^{rd}-, and fourth-order models of the data. The CE values for each model are shown in the second column of Table 1; it can be seen that CE decreases steadily as order increases. Similar experiments to this have been done many times, going back to the classic work of Shannon (1951). What is novel about our approach is the addition of "shuffled" models, shown in the third column of Table 1. Comparing each value in the second column to the value one row down in the third column shows the difference in CE between each nth-order model and the shuffled version of that model created for comparison with the $(n + 1)$-order model. The shuffled values represent the mean of ten iterations of the procedure, as in all the tests reported below. For the 0^{th}, first, and second-order models, the original and shuffled versions are quite close in CE, but for the third-order model they diverge quite markedly (original = 2.113, shuffled = 1.892). Comparing the fourth-order model to the shuffled third-order model in the same row controls for the smaller context counts in the fourth-order model. We can also calculate the percentage improvement in CE of each nth-order model in relation to the shuffled (n–1)-order model; this is how we quantify the predictive benefit of the nth-order model, shown in the fourth column of Table 1.

In what follows, we examine conditional entropy in three melodic corpora. In all cases, the data represents pitch only; rhythm and meter are ignored. Each pitch is encoded with two features: its interval to the previous pitch—a signed integer representing the interval size (in semitones) and direction—and its scale-degree (SD, its position in the scale of the current key), which can be represented as an integer 0 through 11, where 0 is the tonic. Any melody (if the key is known) can be encoded simply as a series of (interval, SD) tuples in this way. This is a very commonly used approach in n-gram models of melody; it is employed in several studies using the IDyOM system (Omigie et al., 2012;

Hansen & Pearce, 2014; Morgan et al., 2019; Sauvé & Pearce, 2019). It can be seen that this encoding is somewhat redundant: Once the scale-degree of one pitch is known, the interval to the following pitch determines its scale-degree. For example, if a note is scale-degree 0 and moves by an interval of +4, the scale-degree of the following note must be 4. Another, logically equivalent, way of thinking about this is to imagine that all melodies are transposed to the same key and that all context-target types are shifted so that the last context note is always within the same octave. We arbitrarily chose the key of C and the octave above (and including) middle C, conventionally represented with the integers 60 through 71. Context and target events can then be represented as pitches: A second-order context might be represented as (64, 67); a target of 55 would represent an interval of −12 to scale-degree 7. For the first note of the melody, there is no preceding interval; we replace it with a special start symbol, "*". Strictly speaking, there is no 0^{th}-order model in our approach, since the contexts are all shifted to end in the same octave. Instead, we simply consider the entropy of (transposed) pitches out of context, which could be considered analogous to a 0^{th}-order model.

Transposing all melodies to the same key is desirable for multiple reasons. By conventional wisdom, musical patterns and principles are invariant across keys: The note C in the context of C major has the same musical properties as the note D in the key of D major. There is an additional reason for transposing melodies in the current situation: the key of a melody affects the sequences of pitches that are likely to occur, and also creates indirect dependencies between pitches. The pattern E4-C4-D4 is much more likely to be followed by E4 than by C#4 or Eb4 (corpus data confirms this), but this is surely due at least partly to the fact that E4-C4-D4 implies a key of C major (or perhaps another closely-related key such as A minor) which makes Eb4 and C#4 unlikely. The effect of key on statistical dependencies between pitches is not of interest in the current study, and it seems best to eliminate it by transposing all melodies to the same key. This also motivates a further step: limiting our study to melodies in major keys. The differences between major and minor scales create further dependencies between pitches. (E4-C4-D4 suggests the key of C major, making E4 more likely than Eb4.) Restricting the data to major-key melodies eliminates this factor. Transposing all melodic contexts to the same octave seems justifiable for similar reasons, since musical patterns and principles are generally viewed as invariant across octaves.

3 Tests

3.1 European Folksongs

Our first test used the Essen Folksong Collection, a corpus of European folksong melodies, a large majority of them German (Schaffrath, 1995). Excluding minor-mode melodies left 5,445 melodies and about 270,000 notes. Table 2 shows the CE of n-gram models from 1^{st}- to 4^{th}-order. As explained earlier, there is no 0^{th}-order model in our framework; in its place, we simply use the entropy of pitches (transposed to C). The value for pitch entropy deserves brief comment. Pitches in the corpus range from G2 to A5—a range of 39 pitches, which would yield entropy of 5.28 if it were uniformly filled. The fact that the entropy is well below this is due to two factors. First, the vast majority of notes are in the middle part of the pitch range. Second, some scale-degrees are more

frequent than others, as is typical of Western tonal music (Krumhansl, 1990): degrees within the major scale (0, 2, 4, 5, 7, 9, and 11) are far more common than "chromatic" degrees, and those within the tonic chord (0, 4, and 7) are especially common.

Table 2. CE of *n*-gram models on the folksong corpus

Model	CE	Shuffled CE of (*n*-1) order	Predictive benefit (%)
Pitch entropy	3.370	-	-
1st-order	2.632	(3.370)	21.9
2nd-order	2.444	2.625	6.9
3rd-order	2.303	2.412	4.5
4th-order	2.110	2.212	4.6

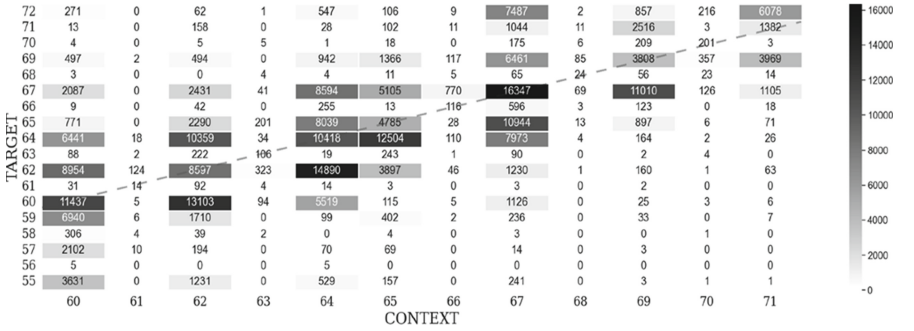

Fig. 2. Frequency of pitch transitions in the folksong corpus.

Not surprisingly, we observe an order effect—a decrease in CE as order increases. As explained earlier, we "shuffle" the contexts at each order so that their sizes match those of the immediately higher order, repeating the process 10 times and taking the mean CE value. We then calculate the predictive benefit (percentage decrease in CE) of each order over the shuffled (*n*–1)th order. In the case of pitch entropy, no shuffling is possible (there is no context), so we simply compare the first-order model to raw pitch entropy.

We examined the data more closely to see how the CE values for different orders—and the differences between them—might be explained. If we compare first-order CE to pitch entropy, this is essentially the question of whether some melodic intervals are more likely than others. It is well established that small melodic intervals are more probable than large ones; the phenomenon of "pitch proximity" has strong support from numerous corpus studies and melodic expectation studies (e.g. Schellenberg, 1997; von Hippel, 2000). The first-order model is represented graphically in Fig. 2, with context pitches on the x-axis and target pitches on the y-axis; essentially, this is a pitch transition table,

with the darkness of each cell representing the frequency of the transition. The dotted diagonal represents repeated pitches. The overall preference for certain scale-degrees can clearly be seen, as well as pitch proximity; each pitch is most likely to move to nearby pitches.

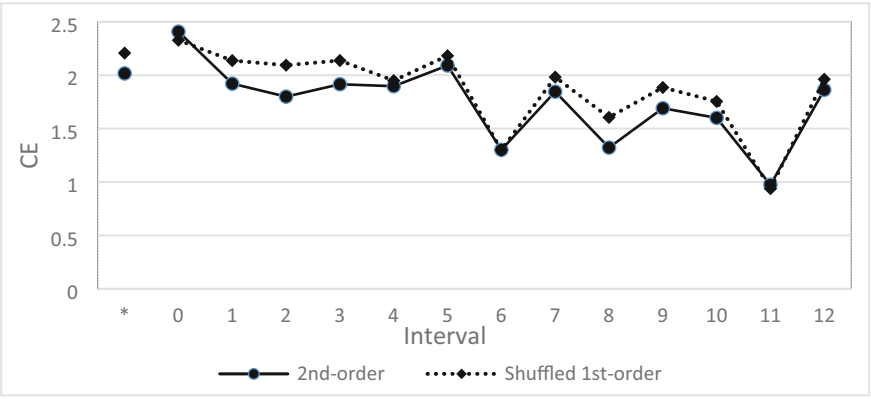

Fig. 3. Second-order CE in the folksong corpus.

We now compare the predictive power of the second-order model to the "shuffled" first-order model. Overall, we see a predictive benefit (a proportional decrease in CE) of 6.9%—markedly less than the predictive benefit of the first-order model (21.9%). To better understand the causes of this predictive benefit, we categorized second-order contexts by the size (absolute value) of the interval between the two context notes (we call this the "context interval") and averaged the CE across contexts for each of these intervals. (These averages are unweighted—giving each context equal weight—as opposed to weighting each context by its frequency, as one would normally do in calculating CE.) These values are shown by the solid line in Fig. 3. Intervals larger than 12 semitones account for less than 0.04% of the data and are not shown here or in later graphs. The leftmost point on the x-axis indicates the case where the target note is the second note of the song. In that case, we define the first symbol of the context as a special "start" symbol, "*"; the second-order context is "* note" instead of "note note." We also calculated the (unweighted) average CE of the shuffled parent contexts of the second-order contexts in each interval category, shown by the dotted line in Fig. 3. The difference between the two values for each interval indicates the predictive benefit of second-order contexts starting with that interval, relative to their first-order parents. An interesting pattern emerges. Second-order contexts beginning with small intervals (1, 2, and 3) have a relatively high predictive benefit over their parents; contexts beginning with quite large intervals (greater than 6) have a similar benefit; contexts with medium-size intervals (4, 5, and 6) have less benefit. Possible reasons for this pattern will be discussed in sect. 4. Intervals 6 and 11 are both rare, possibly causing the CE values of their contexts (and their shuffled parent contexts) to be underestimated; interval 11 is especially rare, with only 20 tokens in the corpus.

Are the differences between intervals shown in Fig. 3 statistically significant? To test this, we compared the predictive benefits (the decrease in CE from first to second order as a proportion of first-order CE) of each pair of intervals, using an unpaired t-test comparing the 10 runs for each interval; see Table 3. The greater predictive benefit of small intervals (1, 2, and 3) over medium-sized ones (4, 5, and 6) emerges clearly. We also see a significantly greater predictive benefit for large intervals than for medium-sized ones, though only for the large intervals 8 and 9.

Table 3. Differences between intervals in predictive benefit*

	1	2	3	4	5	6	7	8	9	10	11	12
0	–	–	–	–	–	0	–	–	–	–	0	–
1		–	0	+	+	+	0	0	0	0	0	0
2			0	+	+	+	+	0	0	0	0	+
3				+	+	+	0	0	0	0	0	0
4					0	0	0	–	–	0	0	0
5						0	0	–	–	0	0	0
6							0	–	–	0	0	0
7								0	0	0	0	0
8									0	0	0	+
9										0	0	0
10											0	0
11												0

*Comparisons between predictive benefits of context intervals in the second-order model (folksong corpus): "+" = row interval is significantly higher in predictive benefit than column interval ($p < .05$ with Bonferroni correction); "–" = significantly lower; "0" = no significant difference.

We now repeat the process just described, but comparing third-order CE to (shuffled) second-order CE. The overall predictive benefit of the increase to third-order shows a further decline, to 4.5%. Again, we categorize third-order contexts by their first interval (see Fig. 4). The resulting pattern is rather different from that observed for the second-order model (shown in Fig. 3). The predictive power that small intervals have with regard to the following interval is much less evident for the interval after that (we will call this the "second following interval"). Predictive benefit increases more or less monotonically as intervals get larger. The correlation between predictive benefit and interval size for intervals 0 through 12 is positive and significant, $r(11) = .58$, $p < .05$. Finally, we compare a fourth-order model to a third-order one using the same procedure as above (see Fig. 5). The trend is very similar to that observed for the second-to-third comparison—predictive

Fig. 4. Third-order CE in the folksong corpus.

benefit increases as intervals get larger. The correlation between interval size (0 through 12) and predictive benefit is again positive, though not significantly so ($r(11) = .30$, n.s.).

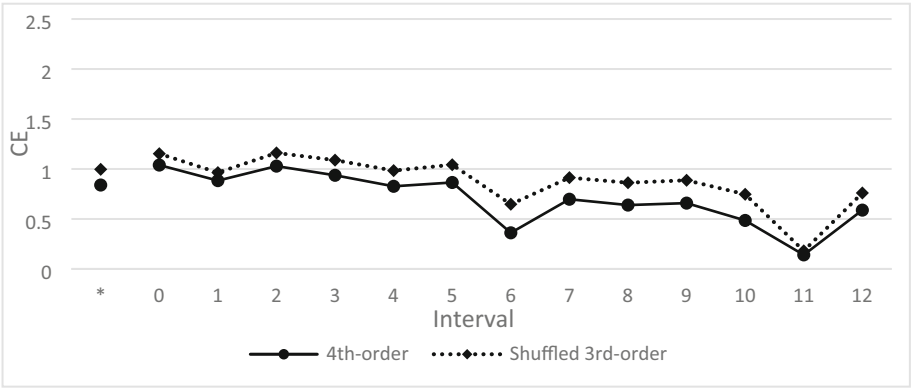

Fig. 5. Fourth-order CE in the folksong corpus.

Table 4. Predictive benefits of first through fourth orders for different corpora

Corpus	0^{th} to 1^{st}	1^{st} to 2^{nd}	2^{nd} to 3^{rd}	3^{rd} to 4^{th}
Text	16.8	17.9	21.4	17.6
Folksongs	21.9	6.9	4.5	4.6
Hymn tunes	37.3	6.8	2.8	1.9
Classical themes	31.5	7.8	4.0	3.1

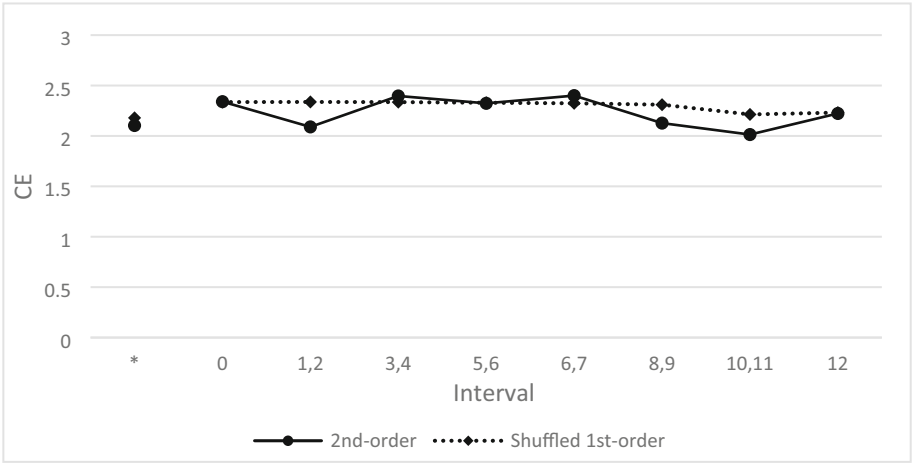

Fig. 6. Second-order CE in the hymn tune corpus. X-axis categories represent diatonic intervals; the numbers show the corresponding chromatic intervals.

3.2 Hymn Tunes and Classical Themes

We now report results of similar tests on two other corpora of melodies. One is a corpus of English-language hymn tunes (N. Temperley, 1998); the corpus contains 17,683 melodies and about 1,050,000 notes. The tunes are encoded using diatonic scale-degree symbols (which we label here as d1 through d7, to avoid confusion): scale-degree 0 is d1; 1 and 2 are both categorized as d2; 3 and 4, as d3; 5 and (sometimes) 6, as d4; 7 and (sometimes) 6, as d5; 8 and 9, as d6; and 10 and 11, as d7. Since notes are already encoded as scale-degrees, no transposition is necessary. The octave of the opening note of each tune is not specified, and subsequent notes are assumed to be in the same octave unless marked with "U" (an octave up) or "D" (an octave down). We arbitrarily mapped the opening octave of each tune on to the diatonic scale-degrees 22 through 28, and encoded subsequent notes in relation to that. Tunes are indicated as being in major or minor, but since both major and minor tunes are represented with the same seven scale-degrees, both types were included. We then followed exactly the same procedure as with the Essen corpus, measuring 0^{th}-order pitch entropy and 1^{st}-, 2^{nd}-, 3^{rd}-, and fourth-order CE, as well as shuffled versions of (n–1)th orders. The results are shown in Table 4. The predictive benefit of first over 0^{th} order is considerably larger for the hymn tunes than for the folksongs, perhaps representing a more constrained use of melodic intervals. For higher orders, however, the predictive benefits are even smaller than those observed with the folksongs. Figure 6 shows the comparison of the second-order model to the shuffled first-order one for each interval. Intervals are reckoned between diatonic scale-degrees and are therefore also diatonic; one category includes chromatic intervals 1 and 2, for example. (Occasionally, a diatonic interval might represent chromatic intervals not shown here, but this is rare.) The pattern is almost exactly like that observed with the folksongs (compare with Fig. 3): predictive benefit is relatively large for small intervals (1 and 2) and for large intervals (8 or larger); it is smaller or even negative for medium-sized intervals. Second-to-third-order and third-to-fourth-order comparisons, not shown

here, reveal very small predictive benefits that do not vary much across intervals. Still, examining intervals up to an octave, there is a trend of increasing predictive benefit for larger intervals (as observed in the folksong corpus), nearly significant for the second-to-third-order comparison ($r(6) = .62$, $p < .1$) and significant for the third-to-fourth-order comparison ($r(6) = .93$, $p < .001$).

Our third test used a corpus of classical instrumental melodies: Barlow and Morgenstern's *Dictionary of Musical Themes* (1948). In order to make the corpus more stylistically homogenous, we limited it to themes dated between 1700 and 1899, and we included only themes in major keys. This yielded 3,789 themes and about 69,000 notes. Notes are encoded in absolute pitch, similar to the folksong corpus, so exactly the same method could be used. Table 4 shows the predictive benefits of each order, which are quite similar to those found for the other melodic corpora. Figure 7 shows the second-order and shuffled first-order models broken down by interval. We find the same qualitative pattern seen in the folksong and hymn corpora, with larger predictive benefits for small and large intervals than for medium-sized ones. At higher orders (not shown here), correlations between predictive benefit and interval size in the classical-theme corpus are in the opposite direction from that observed in the other two corpora: predictive benefit decreases as interval size increases (though not significantly), in both the third-order model, $r(11) = -.38$, n.s., and the fourth-order model, $r(11) = -.17$, n.s. We discuss a possible reason for this below.

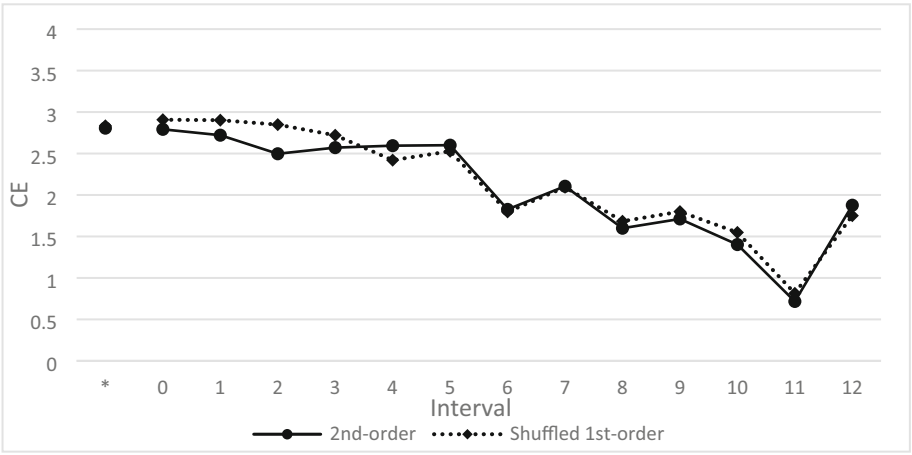

Fig. 7. Second-order CE in the classical theme corpus.

Two further aspects of the interval graphs shown above require discussion. Interval 0 represents a repeated pitch. In all graphs, it can be seen that the predictive benefit of this interval is small: that is to say, preceding a note of the context by another note of the same pitch does not help to predict the following note (or later notes). The leftmost point on each graph represents contexts at the beginning of a melody. For example, in Fig. 3, 1st-order contexts containing a single note are compared to second-order contexts in which that note is preceded by "*", identifying it to be the first note of the melody. It

can be seen that this second-order information has a predictive benefit in the folksong corpus; that is, knowing that a note is the first note of a melody helps to predict what will happen next. For the hymn-tune and classical-theme corpora, this "first-note" effect is very small.

4 Discussion

Our goal in this study was to explore the predictive benefits of order increases in n-gram models on melodic data. We examined the CE (conditional entropy) of different n-gram orders on three musical corpora, using a novel method to control for the fact that CE tends to be underestimated for higher orders. Table 4, presented earlier, gives the crucial results of our three musical corpus analyses, as well as an analysis of text (letters); Fig. 8 shows these results in graphic form (for 1st- through fourth-order models only). In the text corpus, a strong decrease in CE is evident with each increase in order. In the musical corpora, there is a strong predictive benefit of first order over 0th order; no doubt this reflects pitch proximity, the well-known preference for small melodic intervals. The predictive benefits of higher orders are more modest; in Table 4, they range between 1.9% and 7.8%. This can be seen in Fig. 8 as well; the most appropriate comparison is not between adjacent points on a line, but rather, between a point on a solid line and the vertically aligned point on the corresponding dotted line (with the same marker symbol), representing the shuffled lower-order model. In all corpora (text and music), predictive benefits decrease as order increases, reflecting the fact that the predictive power of an element (letter or melodic interval) decreases as its distance from the target element increases.

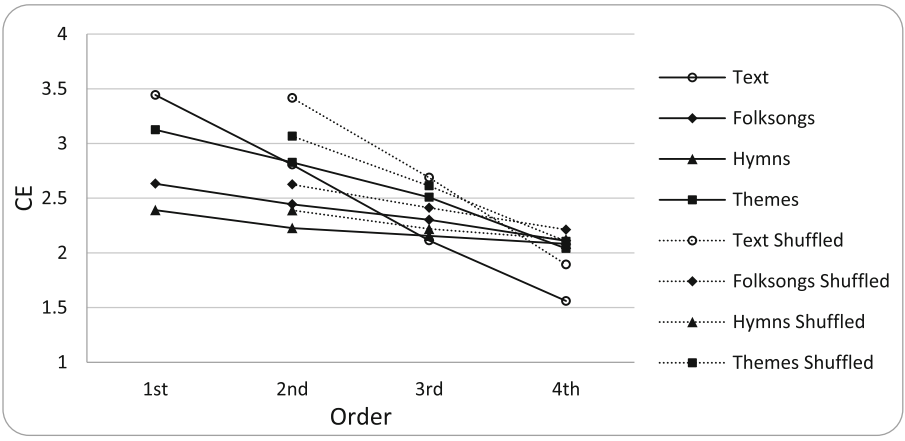

Fig. 8. CE in the four corpora. Each *n*th-order model is vertically aligned with the shuffled (*n*–1)-order model.

We analyzed the 2nd- and higher-order models of the melodic corpora to see how and why their predictive benefits arise. The magnitude of these benefits varies considerably

depending on the first interval of the context. For second-order models, we see more benefit for small and large intervals than for medium-sized ones. This pattern may reflect general principles of melodic structure. Stepwise intervals (half-steps or whole-steps) have a strong tendency to continue with another small interval in the same direction, a phenomenon known as "step inertia" (Huron, 2006; Chiu & Temperley, 2024). After a small interval, then, the following interval is rather predictable. Large intervals have a tendency to be followed by a change of direction, a phenomenon known as "gap-fill" or "reversal" (Narmour, 1990; Schellenberg, 1997). (Narmour [1989] defines a large interval as anything larger than a tritone, i.e. interval 6.) It has also been suggested that gap-fill is an effect of range (Von Hippel & Huron, 2000). If we think of the range of a melody as following a roughly Gaussian distribution (Temperley, 2007), a large leap is likely to span the midpoint of the range, and the following pitch is most likely to move toward that midpoint, causing a change of direction. By contrast, medium-sized intervals do not have any strong directional tendency. Thus, it seems likely that the predictive benefits of the second-order models arise, at least in part, from step inertia and range constraints.

At higher (third and 4th) orders, the advantage in predictive benefit for small intervals disappears, but the advantage for large intervals is still present (though only in the two vocal corpora, not the instrumental one). Again, this pattern of results may be partly due to the effect of range. Imagine a four-note pattern; if the pattern begins with a small interval between notes 1 and 2, this may have little predictive power with regard to note 4. By contrast, a large leap between notes 1 and 2 is likely to span the midpoint of the range; if note 3 stays close to note 2, it is likely that note 4 will move toward the midpoint. This pattern was not found for the classical instrumental themes; in that case large intervals in the 3rd- and fourth-order models do not show greater predictive benefit than small ones. This may be because instrumental melodies are less constrained in terms of range than vocal ones.

In this paper, we have offered, for the first time, an unbiased estimate of the predictive benefits of order increases in melodic n-gram models. Since increases in order incur a substantial increase in complexity (number of parameters), knowledge of these predictive benefits may be useful in optimizing the balance between complexity and accuracy in *n*-gram models. We also show that predictive benefits vary considerably depending on the first interval of the context, in ways that seem to reflect general principles of melodic structure: pitch proximity, step inertia, and range constraints. We have confined our attention to three genres of pre-20th-century Western music: folksongs, classical themes, and hymns. Our approach could also be applied to other styles, such as modern popular music and non-Western styles. It would be interesting to know, for example, whether the high predictive power of contexts beginning with large intervals is found in other styles as well.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

Barlow, H., Morgenstern, S.: A Dictionary of Musical Themes. Crown, New York (1948)

Chiu, M., Temperley, D.: Melodic differences between styles: modeling music with step inertia. Music. Sci. **7**, 1–11 (2024)

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

Goyal, A., Daumé, H., Venkatasubramanian, S.: Streaming for large scale NLP: language modeling. In: Proceedings of Human Language Technologies: the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 512–520 (2009)

Hansen, N.C., Pearce, M.T.: Predictive uncertainty in auditory sequence processing. Front. Psychol. **5**, 1052 (2014)

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

Krumhansl, C.: Cognitive Foundations of Musical Pitch. Oxford University Press, New York (1990)

Morgan, E., Fogel, A., Nair, A., Patel, A.D.: Statistical learning and gestalt-like principles predict melodic expectations. Cognition. **189**, 23–34 (2019)

Narmour, E.: The "genetic code" of melody: cognitive structures generated by the implication-realization model. Contemp. Music. Rev. **4**, 45–63 (1989)

Narmour, E.: The Analysis and Cognition of Basic Melodic Structures. University of Chicago Press, Chicago (1990)

Omigie, D., Pearce, M.T., Stewart, L.: Tracking of pitch probabilities in congenital amusia. Neuropsychologia. **50**, 1483–1493 (2012)

Pearce, M.T.: Statistical learning and probabilistic prediction in music cognition: mechanisms of stylistic enculturation. Ann. N. Y. Acad. Sci. **1423**, 378–395 (2018)

Pearce, M.T., Müllensiefen, D., Wiggins, G.A.: The role of expectation and probabilistic learning in auditory boundary perception: a model comparison. Perception. **39**, 1367–1391 (2010)

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

Pearce, M.T., Wiggins, G.A.: Expectation in melody: the influence of context and learning. Music. Percept. **23**, 377–405 (2006)

Sauvé, S.A., Pearce, M.T.: Information-theoretic modeling of perceived musical complexity. Music. Percept. **37**, 165–178 (2019)

Schaffrath, H.: The Essen folksong collection. In: Ed. by D. Huron. Stanford, CA: Center for Computer–Assisted Research in the Humanities (1995)

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

Schubert, P., Cumming, J.: Another lesson from Lassus: using computers to analyse counterpoint. Early Music. **43**, 577–586 (2015)

Shannon, C.E.: Prediction and entropy of printed English. Bell Syst. Tech. J. **30**, 50–64 (1951)

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

Temperley, N.: The Hymn Tune Index. Clarendon Press, Oxford, UK (1998)

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

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

Youngblood, J.E.: Style as information. J. Music Theor. **2**, 24–35 (1958)