In Press in Cognition

Statistical learning and Gestalt-like principles predict melodic expectations

Emily Morgan[a,b,*], Allison Fogel[a], Anjali Nair[a], Aniruddh D. Patel[a,c,d]

[a] *Department of Psychology, Tufts University, 490 Boston Ave, Medford, MA 02155, United States*

[b] *Department of Linguistics, University of California, Davis*

[c] *Azrieli Program in Brain, Mind, & Consciousness, Canadian Institute for Advanced Research (CIFAR)*

[d] *Radcliffe Institute for Advanced Studies, Harvard University*

Abstract

Expectation, or prediction, has become a major theme in cognitive science. Music offers a powerful system for studying how expectations are formed and deployed in the processing of richly structured sequences that unfold rapidly in time. We ask to what extent expectations about an upcoming note in a melody are driven by two distinct factors: Gestalt-like principles grounded in the auditory system (e.g. a preference for subsequent notes to move in small intervals), and statistical learning of melodic structure. We use multinomial regression modeling to evaluate the predictions of computationally implemented models of melodic expectation against behavioral data from a musical cloze task, in which participants hear a novel melodic opening and are asked to sing the note they expect to come next. We demonstrate that both Gestalt-like principles and statistical learning contribute to listeners' online expectations. In conjunction with results in the domain of language, our results pointing to a larger-than-previously-assumed role for statistical learning in predictive processing across cognitive domains, even in cases that seem potentially governed by a smaller set of theoretically motivated rules. However, we also find that both of the models tested here leave much variance in the human data unexplained, pointing to a need for models of melodic expectation that incorporate underlying

*Corresponding author: *eimorgan@ucdavis.edu*; (530) 754-0994

hierarchical and/or harmonic structure. We propose that our combined behavioral (melodic cloze) and modeling (multinomial regression) approach provides a powerful method for further testing and development of models of melodic expectation.

Keywords: music, melody, expectation, statistical learning, probabilistic modeling

---

1. Introduction

Across cognitive domains, people generate *expectations* or *predictions* about upcoming events (Bubic et al., 2010; Clark, 2013; Friston, 2009). For example, when perceiving complex sequences such as language and music, people predict upcoming words, grammatical structures, notes, chords, etc. (Altmann & Kamide, 1999; DeLong et al., 2005; Huron, 2006; Jackendoff, 1992; Kuperberg & Jaeger, 2015; Levy, 2008; Meyer, 1956; Patel & Morgan, 2016; Rohrmeier & Koelsch, 2012; Tillmann, 2012; Van Berkum et al., 2005; Van Petten & Luka, 2011; Vuust et al., 2009). Such prediction has been hypothesized to contribute to learning (wherein incorrect predictions drive greater learning; Chang et al., 2000, 2006; Dell & Brown, 1991; Fine & Jaeger, 2013; Kidd et al., 2012) and efficient information processing (e.g. aiding understanding speech in noisy environments or accurately reproducing musical rhythms; Clayards et al., 2008; Povel & Essens, 1985). A fundamental question in cognitive science is thus how such expectations are formed—both within a specific domain and across domains.

Here, we focus on the question of expectation in music, specifically *melodic expectations*, or expectations about what notes are coming next in a melody. In music, the ability to form expectations is crucially linked to enjoyment: listeners form expectations about upcoming events, and their enjoyment of the music partly derives from strategically having those expectations confirmed and violated at the right times (Huron, 2006; Jackendoff, 1992; Meyer, 1956). Understanding why humans universally enjoy music thus involves understanding how these expectations are formed.

In the closely related domain of language, accounts of expectation or prediction have demonstrated that predictions rely both on rule-like knowledge and on statistical learning (for example, of n-gram sequences or transition probabilities; Arnon & Snider, 2010; Arnon & Cohen Priva, 2013; Demberg & Keller, 2008; DeLong et al., 2014; Morgan & Levy, 2016; Saffran et al., 1996). The relative importance of these two factors in musical expectations is currently debated (Pearce & Wiggins, 2006; Temperley, 2014). Thus we will focus on comparing theories of melodic expectation that rely on rule-like perceptual principles versus those that rely on statistical learning from one's lifetime experience.

On the one hand, it has been proposed that much like the Gestalt principles that apply in vision (e.g. "good continuation"; Rock & Palmer, 1990), similar rule-like, Gestalt-like principles govern melodic expectations—for example, a preference for subsequent notes to move in small intervals. A key feature of such proposals is that they claim expectations are governed by a small number of relatively simple principles. These principles are not domain-general but are grounded either in music theory or in properties of the auditory system, perhaps stemming from principles used by the auditory system for auditory scene analysis, i.e., segregating auditory 'objects' from complex mixtures of sound (Bregman, 1990; Handel, 1993; Trainor, 2015). Perhaps the best-known example of such a proposal is Narmour's (1989; 1990) Implication-Realization model, which proposes five such principles that are claimed to be innate and universal to music cognition. A more recent example is Temperley's (2008) Probabilistic Model of Melody Perception, which we will describe in more detail in Section 1.1.1.

In contrast, statistical-learning-based models claim that listeners are tracking rich details about the statistics of the input—in particular, the probabilities of n-gram sequences over notes. These theories thus claim that melodic expectation is but one instance of a domain-general statistical learning mechanism, applicable additionally to language acquisition (Cristià et al., 2011; Saffran et al., 1996), adult language processing (Arnon & Snider, 2010; Arnon & Cohen Priva, 2013; Morgan & Levy, 2016), visual sequences and visual scene analysis (Fiser & Aslin, 2016; Kirkham et al., 2002), and the motor system (Schubotz, 2007). While the ability to track n-gram sequences in language and domain-generally is now well established, whether such sequences are used in online music processing is currently less clear. Pearce and colleagues (e.g. Pearce, 2005; Pearce & Wiggins, 2006; Hansen & Pearce, 2014) have proposed that such statistical learning is indeed foundational to melodic expectations and have implemented a framework for learning n-gram models of music known as Information Dynamics Of Music (IDyOM). These models are much richer statistical learning models than the Gestalt-like models: specifying probabilities over many n-gram sequences requires tens of thousands of parameters, orders of magnitude more parameters than required by Gestalt-type models. Because they rely on domain-general learning mechanisms, these statistical learning models explicitly minimize the role of music-theoretically motivated principles and/or principles specific to the auditory system in determining melodic expectations.

This issue of the relative importance of Gestalt-like mechanisms and statistical learning mechanisms in music perception has parallels in other branches of psychology. For example, in theories of art, Arnheim (1969) argued that we have instinctive responses to certain basic visual shapes, which guide our emotional responses to visual art. In contrast, Goodman (1976) argued that our aesthetic response to art is entirely based on learning and sensory experience. This debate has motivated a significant amount of research, which has found that both types of mechanisms are involved in people's aesthetic and emotional responses to art (reviewed in Winner, 2018).

Studying the relative contributions of rule-like principles and statistical learning in forming expectations in music processing also provides an interesting comparison to the study of a similar trade-off in language processing. While music does have culturally-specific rule-like principles (Patel, 2003), musical sequences are more flexible and cannot be said to be strictly "ungrammatical" in the way that language can be. Because musical sequences are not as directly answerable to grammatical "rules," one might a priori expect statistical learning principles to play a relatively greater role in forming expectations in music than in language. Nonetheless, as described above, Arnon & Snider (2010), Arnon & Cohen Priva (2013), Morgan & Levy (2016), and others have argued for a larger-than-previously assumed role of statistical learning of multi-word expressions even in language, which seems potentially more rule-governed. Thus the time seems ripe to look for similar effects in music.

In the remainder of this introduction, we will describe existing computational models of melodic expectation, with a focus on the Temperley and IDyOM models, and discuss what work has previously been done comparing these types of models. In Section 2, we describe an existing behavioral dataset from Fogel et al. (2015) using a novel "musical cloze task," which we will use for our first evaluation of the models. In Section 3 we discuss implementation details of the two models, and in Section 4 we describe how we directly compare these models on the Fogel et al. dataset. In Section 5, we describe a follow-up experiment using a similar task, with convergent findings. Section 6 provides a general discussion and conclusion.

1.1. Computational models of melodic expectation

In the quantitative modeling of music cognition, melodic expectation has been an active and important topic of research for over 20 years (e.g. Eerola et al., 2009, 2002; Krumhansl et al., 1999, 2000; Larson, 2004; Margulis, 2005; Pearce & Wiggins, 2006; Pearce, 2005; Rohrmeier, 2016; Schellenberg, 1997; Sears et al., 2018). Thus a benefit of studying melodic expectation is that there are a number of computationally implemented models reflecting different theories of this phenomenon, which allow us to make precise, testable predictions to compare with empirical human data. Specifically, these models assign probabilities to note sequences. In the formulations used here, two such models will be used to assign probabilities to possible continuation notes given the preceding melodic context. We describe these two models, the Temperley and IDyOM models, in detail.

1.1.1. Temperley model

Temperley's (2008) Probabilistic Model of Melody Perception is a Gestalt-type model, in that it relies on a small number of music-theoretically motivated principles. Specifically, it includes 3 principles:

- The *central pitch tendency* says that "a melody tends to be confined to a fairly limited range of pitches." This is operationalized as a normal distribution over pitches centered around the central pitch for a given melody, which is itself chosen from some normal distribution over pitches (representing the probability of central pitches across melodies).
- The *pitch proximity principle* says that "in general, intervals between adjacent notes in a melody are small." This is operationalized as a normal distribution over pitches centered around the previous note.
- The *key profile* measures "the compatibility of each pitch class with a key," reflecting the fact that certain *scale degrees* (i.e. positions of notes within a scale or key) are known to be more probable than others and to evoke more of a sense of "stability" (Brown et al., 1994; Krumhansl, 1990). This principle is operationalized as the empirical probability (from some training corpus) of each scale degree. (This operationalization is analogous to a Krumhansl key profile, except that the profile is defined by the probability of a note rather than by its stability rating).

These three principles are combined such that the probability of a note is the product of its probabilities under all of these principles, given the context.

Temperley's model is a hallmark Gestalt-type model (Huron, 2006; Krumhansl et al., 2000). Its three principles are interpretable and well attested in music theory. The model makes minimal use of statistical learning (in particular, no note-to-note transitions probabilities or n-grams). It also makes minimal use of harmonic or other hierarchical structure. It does make use of the key of the piece (to determine a note's scale degree for purposes of the key profile), but it does not infer a moment-to-moment harmonic progression, nor does it have any notion of pitch classes or functions (beyond scale degrees), such as, for example, a "leading tone."

In contrast to Narmour's Implication-Realization model, which claims that its Gestalt principles are innate, Temperley remains deliberately agnostic about where the principles come from, noting that the principles may themselves be learned from data.

Empirical support. Temperley (2008) evaluates his model against Narmour's Implication-Realization model on a classic melodic expectation dataset from Cuddy and Lunney. In Cuddy & Lunney (1995), participants heard a two note context and were asked to judge a third note on a 7-point scale from "extremely bad continuation" to "extremely good continuation". Temperley finds that his model outperforms Schellenberg's (1997) state-of-the-art two-factor implementation of Narmour's Implication Realization model, providing a fairly good fit to the rating data ($r = 0.744$), and thus providing some evidence that these Gestalt principles are indeed influencing listeners' expectations.

However, we note that this dataset is potentially a poor test of melodic expectations for a number of reasons. Participants only heard a two note context, and the expectations formed from such an impoverished context may not be representative of expectations in longer melodies. Also, because the rating methodology is cumbersome—participants must hear every possible continuation note in order to judge them—only a small number of context intervals can be tested, and participants heard each context with multiple possible continuation notes over the course of the experiment, potentially confounding their later judgements. We aim to address these limitations in our work.

1.1.2. IDyOM

We will compare Temperley's model with Pearce's (2005) Information Dynamics Of Music (IDyOM) model. IDyOM provides a framework for fitting Markov (i.e. n-gram) models of music. An IDyOM model consists of a probability distribution over every possible note continuation for every possible n-gram context up to a given length.

In addition to learning Markov models over specific pitches, the IDyOM framework can operate on "multiple viewpoints," i.e. it can compute n-gram probabilities over multiple features of the musical surface, including absolute pitch, scale degree, pitch interval from note to note, etc., as well as a limited number of rhythmic viewpoints (e.g. note duration, whether the current note is longer or shorter than the previous, etc.). In our work, we use a "linked viewpoint" of pitch class (i.e. scale degree) and pitch interval between consecutive notes (in semitones)—in other words, our models will learn n-gram probabilities over ordered pairs of (pitch class, pitch interval) or, in IDyOM terminology, (cpint, cpintref). This choice of viewpoints not only follows previous work (Hansen & Pearce, 2014), but also crucially gives IDyOM equivalent information to the information that the Temperley model has, for a fair comparison between the two.

Unlike the Temperley or other Gestalt models, the IDyOM model is a rich statistical learner, in that it stores many n-gram sequences (and hence has many more parameters than the Temperley model). However, it still does not learn any harmonic or other hierarchical structure. (It would in theory be possible to include a harmonic analysis within the IDyOM framework, but such a viewpoint does not currently exist.)

Empirical support. The IDyOM model has also received empirical support as a model of human melodic expectations. Pearce & Wiggins (2006) demonstrate that it outperforms Schellenberg's (1997) two-factor implementation of the I-R model on predicting data from three tasks: the Cuddy and Lunney two-note-context rating task (described above); Schellenberg's (1996) experiment with a similar rating task using eight longer melodic fragments (drawn from British folk songs) as context; and an experiment by Manzara et al. (1992) in which participants provide implicit probability distributions over every note in the melodies of two Bach chorales using a betting paradigm. Pearce et al. (2010) also demonstrated that the IDyOM model can predict neural data including ERP amplitudes and beta band oscillations, while Hansen & Pearce (2014) demonstrated that it can also be used to predict human ratings of uncertainty during music listening. (See also Moldwin et al., 2017, for convergent evidence using a simpler Markov model.)

1.2. Gestalt-like principles versus statistical learning

While both the Temperley and IDyOM models have received empirical support, little work has compared them directly. In particular, this means that we do not know to what extent the predictions made by one could be subsumed by the predictions made by the other. For example, it could be the case that seeming evidence of n-gram learning is actually a case of n-grams capturing specific instances of the general principles embodied by the Gestalt models. Alternately, learning n-grams may in fact be necessary because each n-gram sequence is distinctive, and the Gestalt principles captured by the Temperley model may be post-hoc generalizations drawn by music theoreticians that do not play a true cognitive role. Thus it is important to directly compare the predictions of these two models.

One previous comparison comes from Temperley (2014). In the interest of making the models as comparable as possible, he uses a simplified Markov model (far simpler than IDyOM) and a simplified version of his 2008 model, which he calls the "Gaussian model." The Markov model in particular is simplified in ways that may worsen its predictions relative to IDyOM: it treats scale degree and pitch interval as orthogonal (computing probabilities over them separately and then multiplying them together), rather than treating them as a linked viewpoint (computing probabilities over ordered pairs) as IDyOM can.[1] Temperley also considers unigram, bigram, and trigram models separately, rather than allowing for combinations of these models (known as *interpolation*, a common technique in computational modeling for improving the predictions made by n-gram models). In summary, the Markov model Temperley considers is very

[1] We see in our own data that the linked IDyOM viewpoint frequently—though not always—outperforms the unlinked viewpoint (Appendix A).

simplified relative to IDyOM.

Temperley tests the two models on three tasks: predicting corpus data, predicting the Cuddy and Lunney rating data, and predicting distributions of intervals across melodies. For the third of these (predicting distributions of intervals across melodies) both models do extremely well, providing little basis for useful comparison. For predicting corpus data, the Markov model consistently outperforms the Gaussian model. For predicting human rating data, performance is more mixed but the Markov models generally outperform the Gaussian model. However, Temperley proposes that the actual model performance be weighed against the much larger number of parameters in the Markov model, and hence argues for the Gaussian model on the basis of simplicity.

Given the limitations of the Markov model that Temperley uses, the limitations of the Cuddy and Lunney dataset, and the general inconclusiveness of the results, we think it is well worth revisiting this issue using state of the art models and a richer behavioral dataset. We will return to the issue of comparing number of parameters in the models in the general discussion.

Thus our goal is to do a direct comparison of state-of-the-art versions of both the IDyOM and Temperley models. Moreover, we test them against data that measures expectedness of upcoming notes as directly as possible: we use a "musical cloze" task in which participants hear novel melodic openings and are asked to sing the note or notes that they think should come next. (See Section 2.) We can then compare the empirical probabilities of different notes with the probabilities predicted by the models.

2. Experiment 1: Behavioral Data

We first compare the IDyOM and Temperley models using behavioral data from a new task developed by Fogel et al. (2015). Comparable to a traditional linguistic cloze task, in which participants see the beginning of a novel sentence and are asked to predict what word will come next, participants in the musical cloze task heard the beginning of a novel melody and were asked to "sing the note you think comes next.” Participants found this task easy to do, and we believe it is well suited to reveal participants’ expectations. Moreover, it avoids some problems with traditional tasks that probe melodic expectations such as requiring participants to play the note on a piano (which requires musical training) or asking participants to rate a continuation note following a context (in which all possible continuation notes must be rated, so that contexts are generally heard many times by the same participant in order to collect enough data).

2.1. Materials and Methods

Melodic openings (’melodic stems’) for the task were composed in pairs, such that by changing a small number of notes in the context, Fogel et al. manipulated whether the stem implied an *authentic cadence* (AC condition) or not (Non-Cadence or NC condition; Figure 1). An authentic cadence is a progression from harmony V (a dominant chord), which is subjectively perceived as very unstable and is overwhelmingly followed by harmony I, to the expected harmony I (a tonic chord), producing a sense of resolution; this transition is arguably the most foundational harmonic progression in Western music, and is expected even by non-musically-trained listeners (Loui & Wessel, 2007). Specifically, a melodic stem ending with an implied V harmony would be expected to resolve to a I harmony, and hence participants are expected to sing the tonic (the note with scale degree 1) in the AC condition melodies. NC condition melodies did not end on a V harmony and were designed to not create a strong expectation for any particular continuation note. (There were 45 melodic pairs: any given participant only heard the AC or NC version of a particular melody.) Although the stems used in this task were monophonic melodies, and hence do not contain explicit harmonic material, such melodies still reliably generate implicit harmonic structure for Western listeners (Cuddy et al., 1981; Povel & Jansen, 2002). Melodic stems in each pair were matched for melodic contour, number of notes, rhythm, and key, and averaged 8.4 notes in length. (All melodies are given in Supplementary Materials.) Participants were 50 undergraduates from

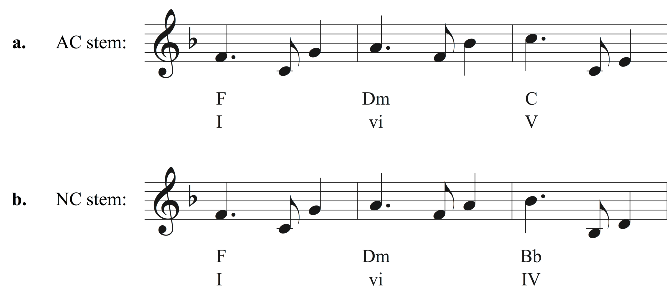

Figure 1: Sample pair of Authentic Cadence (AC) and Non-Cadence (NC) condition melodies annotated with one possible interpretation of the underlying harmonic progression expressed both as chord names (e.g. F, Dm, C) and harmonic functions (I, IV, V).

Tufts University who self-identified as musicians (mean 9 years of formal music training). Full details on the task, stimuli, and participants are available in Fogel et al..

2.2. Preliminary results from Fogel et al.

Figure 2 shows the type of data generated by the melodic cloze task. The left and rights panels show the distributions of responses produced by participants who heard the AC and NC versions of the melodic stems in Figure 1. For the AC stem, the vast majority of the participants sang the tonic. For the NC stem, responses were much more varied. Indeed, across all melodic stems, participants overwhelmingly sing the tonic in the AC condition melodies and not in the NC condition (Figure 3, row 1, Exp 1). We can quantify this difference in multiple ways, including the *constraint* (the probability of the most-commonly sung continuation note, as determined by cloze responses) and the *entropy* of the distribution (an information-theoretic measure of how diffuse responses are). NC condition melodies had substantially lower constraint (41% vs. 69%; $t_{86.90} = 7.74, p < 0.0001$ using a two-tailed unequal variance t-test) and higher entropy (2.27 vs. 1.37; $t_{79.84} = 7.74, p < 0.0001$) of responses than AC condition melodies.

This dataset thus provides a good test for computational models of melodic

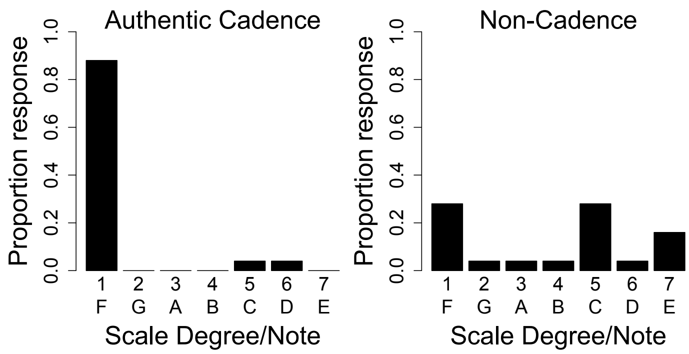

Figure 2: Proportion of responses to sample melody shown in Figure 1. Each column of the histogram shows the proportion of participants who sang that note after hearing the melodic stem in Figure 1a (left histogram) or Figure 1b (right histogram). Individual participants heard one or the other stem.

expectation because it allows us to test (at least) two questions:

1. Can these models can recognize authentic cadences (one of the most important and prevalent instances of expectation in western music)?
2. Can these models make correct diffuse predictions in cases such as the NC melodies where there isn't a single strong expectation?

3. Models

3.1. Training corpora

The Temperley model was original trained on the Essen Folksong Collection (Schaffrath & Huron, 1995). The IDyOM model has been previously been trained on a partially overlapping corpus, which we call the Pearce-Wiggins (PW) corpus (Pearce & Wiggins, 2006; Hansen & Pearce, 2014): The Fink subset of the Essen corpus (consisting of 566 German folksongs), 185 Bach chorale melodies (Bach, 1892; Center for Computer Assisted Research in the Humanities, 1994), and 152 Nova Scotian songs (Creighton, 1966; Sapp, 2018). While the Essen corpus is larger, the Pearce-Wiggins corpus is more stylistically diverse—in particular, it contains composed melodies as well as folk songs. We report results training both models on both possible corpora. (In order to do an apples-to-apples comparison, we always report comparisons in which the models to be compared are trained on the same corpus.)

We next describe the details of how we trained and made predictions from our two models of interest.

3.2. Temperley model

The parameters required to specify the Temperley model are: the mean and variance of the central pitch profile, the variances of the range and pitch proximity profiles, major and minor key profiles, and probability of a major versus minor key. All these parameters can be computed straightforwardly from a training corpus. Temperley kindly provided us with code to run this model, with parameters calculated from the Essen Folksong Collection (as reported in his 2014 paper). We additionally computed the parameters from the Pearce-Wiggins corpus in order to run a version of the model trained on that corpus.[2] For purposes of computing the key profiles from the PW corpus, we assumed all pieces were in a major key (which was consistent with the high probability of a major and not a minor third in the resulting key profiles). All test melodies were in major keys, so it was not necessary to compute a minor key profile from the PW corpus.

In its original formulation, Temperley's model is a Bayesian model in that it computes the probability of an upcoming pitch given the musical prefix for all possible keys, and then marginalizes over keys to get the probability of the target note. We modified the model to report the probability of a continuation note *given the key of piece*, rather than marginalizing over keys. We did this for two reasons: First, doing so makes the Temperley model more comparable to the IDyOM model, which is also given the key of piece (as is required to translate

[2]The pieces in the Pearce-Wiggins corpus were not annotated with their mode, so we could not compute the probability of a major versus minor key from this corpus. However, as described below, we used a modified version of the Temperley model which was given the correct key of each test melody, so this parameter wasn't necessary.

pitches into scale degrees). Second, initial tests showed that the Temperley model performed better as predictor of human data when given the key versus when marginalizing over keys, so using the key-given version gave this model the best chance to perform well in comparison with the IDyOM model.

For the Temperley model trained under each corpus, we computed for all our test melodies the probability of all possible continuation notes from midi note 47 to 83 (aka. B2 to B5).

3.3. IDyOM model

The IDyOM model is publicly available (Pearce, 2005). We trained it using a linked viewpoint of pitch class and pitch interval between consecutive notes, or (cpint cpintref). Both the long term and short term models were used. As with the Temperley model, the long term model was trained on both the Essen and Pearce-Wiggins corpuses. All other model parameters were left as defaults. Again, using the model trained under each corpus, we computed for all our test melodies the probability of all possible continuation notes from midi note 47 to 83 (aka. B2 to B5).

4. Experiment 1: Model Evaluation and Results

4.1. Initial visualization of model predictions

We begin by visually inspecting the predictions made by both the Temperly and IDyOM models (Fig. 3). The first striking thing we notice is that both models severely underpredict tonic responses (i.e. scale degree 1) in the AC condition (and in turn predict much more diffuse responses across other scale degrees). In other words, both models are underconfident in recognizing the implied authentic cadence. This suggests that the answer to Question 1 in Section 2 (Can these models recognize authentic cadences?) is No, or at least that the models are underconfident in their recognition of such cadences. This suggests that there is a need for implicit harmonic structure to be explicitly represented in these computational models, even if the models' aim is only to predict the melody and not the harmony. (See Arthur, 2017; Kim et al., 2018, for convergent evidence.)

We also conclude from this that when we continue with further analyses, we should pay particular attention to the NC conditions. We already know that both models are doing a relatively poor job of predicting responses in the AC condition, but (given their lack of explicit harmonic representations) they may do better in cases where cadence-based expectations are not at play.

4.2. Model comparison

We use multinomial discrete-choice logit modeling (Agresti, 2002) to evaluate the predictive power of both the Temperley and IDyOM models as predictors of the human behavioral data. Multinomial logit models are a generalization of logistic regression which predict the probability of choosing between some number (more than two) of categorical outcomes—in this case, continuation notes. In discrete choice logit models, the value of the predictor can depend on the outcome (e.g. in this case, the value of the Temperley and IDyOM models that we use as predictors depends on the outcome note, as opposed to predictors like subject age that would be constant across outcomes). The `mlogit` package in R (Croissant, 2013; R Core Team, 2016) allows us to fit such models with by-subjects random effects. Specifically, these models allow us to determine what combination of the independent/predictor variables (IVs) best predict the dependent/outcome variable (DV). Crucially, a statistically significant effect of one IV implies that this IV has predictive power above and beyond what is being explained by the other IVs—i.e. it accounts for a statistically significant amount of unique variance. The use of these models thus allows us to test whether the Temperley and IDyOM models are explaining the same or unique variance in the human data.

For each test dataset and training corpus of interest, we fit the model:

$$\texttt{human} \sim \texttt{Temperley + IDyOM}$$

The model does not include a fixed-effect intercept. We additionally include

		Coeff. Estimate	*Std. Error*	*t value*	*p value*
Essen	Temperley	0.44	0.036	12.46	$< 2 \times 10^{-16***}$
	IDyOM	0.61	0.028	21.49	$< 2 \times 10^{-16***}$
PW	Temperley	0.43	0.034	12.77	$< 2 \times 10^{-16***}$
	IDyOM	0.60	0.030	20.15	$< 2 \times 10^{-16***}$

Table 1: Experiment 1: Model fit for all data (AC+NC conditions), with models trained on either Essen or Pearce-Wiggins corpus. The table shows regression coefficients, as well as standard errors, t, and p values, for both model predictors. * indicates statistical significance.

by-subject intercepts and random slopes of both the Temperley and IDyOM predictors. We run this model comparison for Temperley and IDyOM predictors trained on both training corpora, and using the whole dataset (AC+NC) as well as using just the NC subset of data.

All variables (IVs and DV) are coded as scale degrees, collapsing across octaves. In other words, for the human data, the outcome (DV) is coded as the scale degree that was sung, regardless of octave. For the model predictions, for a given melody, we add up the model's predictions for a given scale degree across all octaves, and use the log of this probability as the IV for that scale degree. (See Supplementary Materials for graphs of human data and model predictions which include octave information.) In order to have a tractable number of outcome categories, we consider only in-key notes. (Out-of-key notes were sung 5.3% of the time in the human data in the AC condition and 9.7% in the NC condition, and in many cases were likely instances of poor singing in which the participant intended to sing an in-key note. For example, 51% of out-of-key notes in the AC condition were the minor second and were likely intended to be the tonic.)

4.3. Results

As seen in Tables 1 and 2, both the IDyOM and the Temperley models are significant predictors of human data, across both training corpora and data subsets, suggesting that both statistical learning and Gestalt-like principles make

		Coeff. Estimate	*Std. Error*	*t value*	*p value*
Essen	Temperley	0.46	0.050	9.22	$< 2 \times 10^{-16***}$
	IDyOM	0.35	0.037	9.55	$< 2 \times 10^{-16***}$
PW	Temperley	0.31	0.046	6.75	$1.44 \times 10^{-11***}$
	IDyOM	0.43	0.036	12.00	$< 2 \times 10^{-16***}$

Table 2: Experiment 1: Model fit for NC data only, with models trained on either Essen or Pearce-Wiggins corpus. The table shows regression coefficients, as well as standard errors, t, and p values, for both model predictors. * indicates statistical significance.

independent contributions to human melodic expectations. Note that because the model predictors are both log probabilities, and hence measured on the same scale, we can compare the coefficient estimates directly. Looking at all model fits (both training corpora and both data subsets), IDyOM generally outperforms the Temperley model, as measured in larger coefficient estimates and t values, and smaller p values. This suggests that statistical learning may play a slightly larger role than Gestalt principles in determining human expectations.

4.4. Error analysis

Both the IDyOM and Temperley models leave much variance in the human data unexplained. To quantify this, we define an error metric for each melody (under a given model) by taking the absolute value of the difference in probability between human responses and the model prediction for each possible continuation note, summing these values, and dividing by two. This gives a number between 0 and 1 representing the amount of probability mass that would need to be moved in order to turn one distribution into the other (where higher numbers = more error). Mean errors for each model are reported in Table 3. For example, for the Essen-trained IDyOM model, the mean error is 0.48 for AC and 0.46 for NC melodies. In other words, the model is putting barely more than half of the probability mass in the right place. Looking at individual melodies to see how model predictions differ from the human data may lead to insight about further factors that influence human melodic expectations. We

		AC	NC
Essen	Temperley	0.56	0.45
	IDyOM	0.48	0.46
PW	Temperley	0.54	0.46
	IDyOM	0.52	0.44

Table 3: Average melody-level error (see Section 4.4) for each model (Experiment 1)

have included error measures for all melodies, as well as graphs of the human data and the model predictions for all melodies, in Supplementary Materials.

For example, one melody for which human and model predictions interestingly diverge is NC43 (shown in Figure 4). A substantial proportion of human participants continued this melody with B♭3, which is unpredicted by any of the models (see Figure 5). This effect in the human data likely arises from "stream segregation" (Huron, 2001) wherein the large intervals between successive pitches in the melody, contrasted with the stepwise motion of every other pitch, cause the lower notes (in particular, D4 and C4 in the last two measures) to be perceived as a separate melodic line from the higher notes (B♭4 and A♭4). B♭3 is a natural continuation of the stepwise motion of the D4-C4 sequence, but goes unpredicted by models that cannot separate the lower stream from the higher stream. We believe this represents another instance of the need for hierarchical structure in models of melodic expectation: hierarchical structure is not purely used to represent harmony but is also necessary to represent other aspects of the way melody itself is perceived.

We further notice that even among the NC melodies, some of melodies on which the models perform worst are those in which many participants sing the tonic (e.g. melodies NC14 and NC44; see Supplementary Materials). We previously pointed out that both models underpredict the tonic for AC melodies, but it also worth noting that the IDyOM model underpredicts tonic responses in the NC condition (Fig. 3). This could imply that human expectations are systematically biased towards the tonic, even beyond its true distribution in corpus data. (For a similar comparison case, Huron, 2006, demonstrates that in skip-reversal patterns, trained musicians expect the reversal after a skip beyond what is justified by the statistics of the input.[3])

Another possibility is that the tonic responses we see in the human data could be influenced by task-specific demands. In particular, although participants in the cloze task were instructed to "continue but not necessarily complete" the phrase, they may nonetheless have been biased to find a continuation note that provided a sense of closure. If so, this would be a major confound in our results. To rule out this possibility, we ran a follow-up experiment in which participants were allowed to sing as many notes as necessary to complete the phrase.

5. Experiment 2

Experiment 2 was identical to Experiment 1 except that participants were instructed to "complete [the melody] by singing up to a few notes". 50 self-identified musicians (26 female, 24 male; age range 18-26, mean age 21) with 5+ years of musical experience in the last 10 years participated in the experiment. Participants had an average of 9 years (sd 5 years) of formal musical training. 72% reported "voice" as one of their instruments. Participants were compensated for their participation. Materials were identical to those used in

[3]A skip-reversal is a common pattern in Western music wherein a large leap in pitch (a 'skip') is followed by movement in the opposite direction (the 'reversal'), e.g. a large ascending interval would commonly be followed by a descending interval. Von Hippel & Huron (2000) demonstrated via musical corpus statistics that this pattern is entirely predicted by the general phenomenon of regression to the mean, and therefore its prevalence in corpus statistics requires no special explanation in terms of either physical or cognitive properties of music. Nonetheless, Huron (2006) further found that trained-musician listeners, after hearing a skip, expect to hear a reversal even more strongly than is justified by the regression to the mean phenomenon, and indeed even more strongly than is justified by the statistics of reversals following skips in musical corpora. He concludes that while the skip-reversal pattern may initially have arisen merely from regression to the mean, trained musicians have nonetheless extracted it as a known pattern from their musical experience and/or training, such that they now expect to hear it out of proportion to how frequently it in fact occurs.

		Coeff. Estimate	*Std. Error*	*t value*	*p value*
Essen	Temperley	0.29	0.034	8.55	$< 2 \times 10^{-16}$***
	IDyOM	0.57	0.026	21.65	$< 2 \times 10^{-16}$***
PW	Temperley	0.31	0.033	9.50	$< 2 \times 10^{-16}$***
	IDyOM	0.51	0.028	18.59	$< 2 \times 10^{-16}$***

Table 4: Experiment 2: Model fit for all data (AC+NC conditions), with models trained on either Essen or Pearce-Wiggins corpus. The table shows regression coefficients, as well as standard errors, t, and p values, for both model predictors. * indicates statistical significance.

Experiment 1.

5.1. Behavioral results

The revised task successfully elicited multi-note continuations. Participants sang an average of 4.06 notes (sd 1.05) for AC melodies and 4.73 notes (sd 1.00) for NC melodies. Participants sang a one note completion on 31.0% of AC condition trials and 13.5% of NC condition trials.

Because our computational models specifically make predictions about the next note in a melody, and for direct comparison with the results from Experiment 1, we analyze only the first note in each continuation. These data are shown (aggregated across melodies) in Figure 3, row 1 (Exp 2), and for each individual melody in graphs in the Supplementary Materials. Looking both at Figure 1 and at the individual melody graphs, we notice a striking convergence in the results between Experiments 1 and 2, suggesting that the results of Experiment 1 were not substantially biased by a task-specific tendency to find a single note that would provide a sense of closure. To confirm this impression, we rerun the computational model comparisons using the Experiment 2 data.

5.2. Model comparisons

We begin by noting that the computational models predict the next note without regard to whether it is the final note in a melody or not. Thus, the

		Coeff. Estimate	*Std. Error*	*t value*	*p value*
Essen	Temperley	0.31	0.047	6.60	$4.07 \times 10^{-11***}$
	IDyOM	0.35	0.035	9.93	$< 2 \times 10^{-16***}$
PW	Temperley	0.21	0.046	4.61	$4.02 \times 10^{-6***}$
	IDyOM	0.40	0.037	10.79	$< 2 \times 10^{-16***}$

Table 5: Experiment 2: Model fit for NC data only, with models trained on either Essen or Pearce-Wiggins corpus. The table shows regression coefficients, as well as standard errors, t, and p values, for both model predictors. * indicates statistical significance.

model predictions are identical for Experiments 1 and 2. We run the same multinomial discrete-choice logit analyses for Experiment 2 as we did for Experiment 1. Results are shown in Tables 4 and 5. We again find that across both data subsets and both training corpora, both the IDyOM and the Temperley models are significant predictors of human data, again suggesting that both statistical learning and Gestalt-like principles make independent contributions to human melodic expectations. We again find that the IDyOM model slightly outperforms the Temperley model, as measured in larger coefficient estimates and t values, and smaller p values.

The results of Experiment 2 are thus entirely consistent with those of Experiment 1, implying that the results of Experiment 1 are not due to a bias to sing a note that provides a sense of closure in the single-note-continuation task. In melodies that end with an implied Authentic Cadence, participants overwhelmingly sing the tonic even when it is not the final note they will sing, but these tonic responses are severely underpredicted by both models. Moreover, as described in Section 4.4, participants also sing the tonic in response to NC melodies more so than is predicted by the IDyOM model (though the Temperley model does better in this regard), suggesting that melodic expectations are biased towards the tonic over and above the extent to which it is justified by the statistics of the input.

6. General Discussion

We set out to investigate whether melodic expectations stem from rule-like Gestalt principles or from statistical learning. Specifically, we asked to what extent two state-of-the-art computational models of melodic expectation—Temperley's Probability Model of Music Perception and Pearce's IDyOM model—predict human responses in a musical cloze task. In two experiments, we demonstrated that both models contribute significantly and independently to predicting the human data, suggesting that both Gestalt principles and statistical learning contribute to human expectations. Across all ways of analyzing the data, the IDyOM model tended to be a stronger predictor of the behavioral data, suggesting that expectations rely somewhat more heavily on statistical learning than Gestalt principles. In other words, we conclude that listeners track the probabilities of n-grams of notes over the course of their lifetime of musical experience, *and* that they are sensitive to simple music-theoretically motivated, Gestalt-like principles, and that both of these knowledge sources play a role in shaping expectations for upcoming notes.

We additionally showed that both models failed to recognize authentic cadences, underpredicting responses of the tonic in cases where participants sang that note overwhelmingly. We conclude that implicit harmonic structure plays an important role—not currently recognized by either model—in determining human melodic expectations. Other types of hierarchical structure such as an ability to segregate melodic streams (see Section 4.4) also likely play a role in human melodic expectations, and again are not captured by either of the models considered here.

Our current investigation used musically trained participants, raising the question of whether our results would generalize to non-musically-trained individuals. Our prediction is, broadly speaking, that our findings would hold in non-musically-trained individuals as well. Individuals without musical training are known to form expectations about both melody and harmony, although the ability to attend to multiple aspects of music (such as melody and harmony) simultaneously may be strengthened by musical training (Bigand et al., 2000; Bigand & Poulin-Charronnat, 2006; Koelsch et al., 2002; Loui & Wessel, 2007; Tillmann, 2012). Indeed, the ability to form these expectations in music is thought to be fundamental to the enjoyment of music, a phenomenon which is certainly not limited to trained musicians (Huron, 2006; Meyer, 1956). We know of no theoretical reason why non-trained individuals should not have access to both Gestalt-type principles and statistical knowledge, noting that all individuals growing up in a Western culture will have significant, regular exposure to music, even without formal training. Of course, future work could test our prediction by repeating the experiments presented here using participants without musical training.

6.1. The role of simplicity in evaluating theories

Our work builds on the somewhat-mixed results of Temperley (2014), who found that Markov models generally out-performed his Gaussian model on a variety of tasks. However, Temperley argued for his Gaussian model on the grounds of simplicity, specifically highlighting that it requires far fewer parameters than any Markov model. While we agree that favoring a simpler hypothesis is a useful heuristic, we argue it cannot take the place of or overcome empirical data that actually favors one hypothesis over the other. Here we have presented empirical evidence that human expectations indeed rely on knowledge of n-gram probabilities that cannot be abstracted into the Gestalt principles of Temperley's (2008) model, but that they likewise rely on Gestalt-like principles which are not captured by n-gram probabilities.

We also note that the number of parameters in a computational model is not the only possible measure of simplicity, particularly when we view theories of melodic expectation within the context of other theories of cognition. For example, in language processing, tracking the probabilities of multi-word expressions (comparable to tracking statistics of multi-note n-gram sequences in music) was once thought to be infeasible for human learners due to memory limitations (Pinker, 2000). But we now know that probabilities of even fairly low frequency multi-word expressions are indeed stored and used in online language processing (Arnon & Snider, 2010; Arnon & Cohen Priva, 2013; Morgan, 2016). Given how many more words there are than musical notes (as an approximation, there are 88 keys on a piano, which already spans a much larger pitch range than is typically encountered), to suggest that we further store note n-gram probabilities seems relatively little burden compared to the number of word n-gram probabilities we already know are stored. Indeed, given our knowledge that word n-gram probabilities *are* stored, and given the similarities between the two domains, it could be argued that the simplest theory from a broader cognitive perspective is that note n-gram probabilities would also be stored.

6.2. Cognitive models combining statistical learning and Gestalt-like principles

Our finding that both statistical learning and Gestalt-like principles influence melodic expectations raises a new question: what sort of cognitive process might combine these two types of knowledge in determining melodic expectations? Broadly speaking, we envision two possible types of answers: in one case, statistical learning and Gestalt-like principles operate independently, and then their predictions are combined. In the other case, these two types of principles might in fact emerge from a single system.

In the first case, two types of expectation might come about roughly as their current proponents have suggested: a small set of principles specific to the auditory domain generates one set of expectations, while a domain-general statistical learning mechanism generates another, and these two sets of expectations are combined in some weighted fashion to determine online expectations during music listening, to generate responses in the musical cloze task, etc. While the multinomial logit models we use for data analysis are not designed to be cognitive models, we will note that this type of weighted combination is exactly what they do, providing an algorithmic proof of concept for this method of combining expectations.

In the second case, a single system might be capable of learning both types of knowledge. For example, recent research on Gestalt principles of vision suggests that they may be rational solutions to a statistical inference problem, rather than needing to assume that these principles are specified a priori (Froyen et al., 2015). Indeed, Temperley himself points out that his model's principles might be learned from the input. However, learning the conceptual structure of a system is potentially a much more difficult task than learning the correct values of known parameters. For example, for Temperley's model to be learned via the statistics of the input, the learner would not only need to learn the correct value of the mean and variance parameters for e.g. the central pitch profile, but would need to learn that the central pitch tendency itself is the correct principle to follow in the first place (as opposed to a uniform distribution over pitches in a given range, a disjoint set of possible pitch ranges, or any of infinitely more possible pitch distributions). At least from our perspective as cognitive scientists, this seems like a much more difficult to problem to model a solution for. We know of no proposals for how this might be solved in the domain of music. But, on the other hand, humans clearly are capable of doing this type of abstract reasoning/conceptual structure learning in general, as it seems to be necessary for understanding complex real-world situations (and thus understanding how humans can do this in general is an important question for cognitive science; Kemp & Tenenbaum, 2008). Indeed, there is some evidence from computational models that it is beneficial to simultaneously learn the conceptual structure of a domain along with the values of particular parameters, and the models that do so can take advantage of both domain-specific knowledge and of domain-general statistical learning mechanisms (Tenenbaum et al., 2006). Such an approach may also prove fruitful for modeling how people could learn to generate melodic expectations from n-grams and from Gestalt-like principles simultaneously. (However, we caution that the current examples of such models use highly simplified situations, and so a fully implemented model of melodic expectation along these principles may not be available in the near future!)

6.3. Inferring probabilities from the cloze task

It is important to ask to what extent the responses provided by participants in the cloze task (and the resulting probability distribution over notes) accurately reflect their subjective probabilistic beliefs about upcoming notes. Our implicit assumption in this work has been that participants sample from their subjective probability distributions to generate their outputs in the production task. However, in the linguistic cloze task, Staub et al. (2015) have suggested that the distribution of cloze responses more likely reflects the effects of different levels of activation of word candidates as implemented in a race model, rather than a direct sample from participants' subjective probability distributions.

While the cloze distribution may not exactly reflect a sample from participants' subjective probability distribution, it also might not be far off. In general, we know that cloze probabilities are a strong predictor of human data (both behavioral and neural) in language tasks (e.g. DeLong et al., 2005; Rayner & Well, 1996). Moreover, in language, cloze responses are actually a better predictor of reading times than true corpus probabilities, suggesting that cloze responses are tracking something truthful about subjective probabilities beyond what is realized in the corpus data (Smith & Levy, 2011).

Ultimately, while recognizing that the cloze responses might not provide a perfect mirror of subjective probabilities, we nonetheless consider cloze data at least as good a way of tapping into these subjective probabilities as a more traditional rating task, in which the mapping from ratings to subjective probabilities is entirely unclear. Of course, future work could attempt to replicate the results using a variety of methodologies, including rating tasks as well as neuroscientific methods (discussed further in Section 6.4). In the idealized future, a full theory of melodic expectation would not only capture true subjective probabilities but also, to the extent that these probabilities may appear to differ as a function of the task used to elicit them, would explain what cognitive processes cause these differences in mapping between subjective probabilities and the behavioral/neuroscientific results.

6.4. Future work

We believe that the combination of the musical cloze task with the use of multinomial regression to directly compare models represents a productive and powerful approach to testing future theories of melodic expectation. Any implemented computational model of melodic expectation (which can make predictions about upcoming notes given a musical context) can be tested via this approach. For example, in future work we would like to compare the current models against models that include harmonic structure (Margulis, 2005; Rohrmeier, 2011) or that take rhythmic information into account (van der Weij et al., 2017). We can also develop musical cloze stimuli to probe other facets of melodic expectation, such as other types of cadences or the interaction of melodic expectation with rhythmic prediction (e.g. do listeners form different melodic expectations for stronger versus weaker beats in the metrical hierarchy?). In fact, the modeling and cloze paradigms can work hand-in-hand: we can use computational models to identify moments in music (either from existing musical corpora or in constructed stimulus materials) where different models' predictions diverge, potentially pointing to musical phenomena that are diagnostic of the different predictions made by different theories. We can then test these moments specifically using the cloze paradigm, and finally compare the model predictions to the human data using regression modeling as we did here. (We note in passing that the rise of internet-based auditory testing may permit the collection of large melodic cloze datasets relatively quickly, using new methods that ensure participants are wearing headphones, and automated pitch tracking algorithms to measure sung responses; Woods et al., 2017)

Another direction for future research is to use discrepancies between model predictions and human expectations (in our dataset or others) to develop ideas for new principles to incorporate in models of melodic expectation. For example, as described in Section 4.4, we have identified some melodies in which, even in the NC condition, participants tend to sing the tonic more than predicted by the IDyOM model, potentially pointing to a need to incorporate a specific bias towards predicting the tonic. We also discussed a case of stream segregation that neither model can capture, pointing to a need for hierarchical structures to represent separate melodic streams. In the supplementary materials we have provided our experimental items (both in music notation and audio format). For each melody, we also provide histograms depicting human responses and model predictions, and each model's error (as defined in Section 4.4). We hope this may be of use to researchers searching for principles lacking in current state-of-the art models of melodic expectancy.

Finally, we feel that combining the melodic cloze and current modeling approach with neuroscientific methods could provide a rich area for exploration. Specifically, neuroimaging experiments with stimuli from a melodic cloze study (such as Fogel et al. 2015) can be designed to precisely engineer the degree of melodic 'surprise' of a given note following a given stem. Using such controlled stimuli, the strength of a neural response to an unexpected note in auditory cortex can be quantitatively compared to its probability according to either human melodic cloze data or a computational model of melodic expectation. We can ask which is a better predictor of the amplitude of the neural response: the probability of the note according to melodic cloze measurements, or its probability according to a model of melodic expectation (such as IDyOM)? Initially one might think that probabilities based on cloze data should be a better predictor, since such data are based on human expectations. Yet, as discussed in Section 6.3, the probabilities of notes sung in a melodic cloze paradigm may not be a simple linear reflection of underlying probabilities of tone sequences as tracked by the auditory cortex. Combining data from auditory cortical responses, behavioral paradigms, and statistical learning models such as IDyOM might better allow us to triangulate any non-linear relationships between these phenomena (Pearce et al., 2010; Hsu et al., 2015). More generally, we feel that combining the melodic cloze paradigm with computational models of expectation and neuroimaging methods can provide a powerful new way to study the cognitive science of predictive processing in music.

Model	*All data*	*NC data*
Temperley (Essen)	-3725.1	-1860.0
Temperley (PW)	-3681.5	-1865.0
IDyOM (Essen, linked viewpoint)	-3394.9	-1820.6
IDyOM (PW, linked viewpoint)	-3448.4	-1786.0
IDyOM (Essen, unlinked viewpoints)	-3518.6	-1769.4
IDyOM (PW, unlinked viewpoints)	-3616.3	-1792.3

Table A.6: Log-likelihood of individual model fits as described in Appendix A. Larger (i.e. less negative) values indicate better fit..

Acknowledgements

This work has benefited from presentation and discussion at a variety of venues, including the 2017 Society for Music Perception and Cognition Meeting. Our sincere thanks to David Temperley for many helpful discussions and for providing us with the code for his model.

Declarations of interest: none.

Appendix A. Individual model performance

For every melodic expectation model under consideration, we entered the model's predictions as the sole predictor variable in a multinomial discrete-choice logit model (as described in Section 4.2), for both the whole dataset and the NC melodies only as dependent variables. To show the relative performance of all models, we report the log-likelihood of each model fit (Table A.6).

References

Agresti, A. (2002). *Categorical Data Analysis*. Wiley-Interscience.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, *73*, 247–264.

Arnheim, R. (1969). Art and Visual Perception: A Psychology of the Creative Eye. Revised edition.. Berkeley: University of California Press.

Arnon, I., & Cohen Priva, U. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*, *56*, 349–371.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*, 67–82.

Arthur, C. (2017). Taking harmony into account. *Music Perception: An Interdisciplinary Journal*, *34*, 405–423.

Bach, J. S. (1892). *Bach-Gesellschaft Ausgabe, Band 39* . Leipzig: Breitkopf & Hrtel.

Bigand, E., McAdams, S., & Forêt, S. (2000). Divided attention in music. *International Journal of Psychology*, *35*, 270–278.

Bigand, E., & Poulin-Charronnat, B. (2006). Are we "experienced listeners"? A review of the musical capacities that do not depend on formal musical training. *Cognition*, *100*, 100–130.

Bregman, A. S. (1990). *Auditory Scene Analysis*. The Perceptual Organization of Sound. The MIT Press.

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

Bubic, A., Yves von Cramon, D., & Schubotz, R. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, *4*, 1–15.

Center for Computer Assisted Research in the Humanities (1994). Chorales, BWV 253-438.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*, 234–272.

Chang, F., Dell, G. S., Bock, K., & Griffin, Z. M. (2000). Structural priming as implicit learning: A comparison of models of sentence production. *Journal of Psycholinguistic Research*, *29*, 217–230.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181–204.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*, 804–809.

Creighton, H. (1966). *Songs and Ballads from Nova Scotia*. New York: Dover.

Cristià, A., McGuire, G. L., Seidl, A., & Francis, A. L. (2011). Effects of the distribution of acoustic cues on infants' perception of sibilants. *Journal of Phonetics*, *39*, 388–402.

Croissant, Y. (2013). *mlogit: multinomial logit model*. URL: `https://CRAN.R-project.org/package=mlogit` r package version 0.2-4.

Cuddy, L. L., Cohen, A. J., & Mewhort, D. J. K. (1981). Perception of structure in short melodic sequences. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 869–883.

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

Dell, G. S., & Brown, P. M. (1991). Mechanisms for listener-adaptation in language production. In D. J. Napoli, & J. A. Kegl (Eds.), *Bridges Between Psychology and Linguistics* (pp. 105–129). Hillsdale, NJ.

DeLong, K. A., Troyer, M., & Kutas, M. (2014). Pre-processing in sentence comprehension: Sensitivity to likely upcoming meaning and structure. *Language and Linguistics Compass*, *8*, 631–645.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*, 1117–1121.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*, 193–210.

Eerola, T., Krumhansl, C. L., & Toiviainen, P. (2002). Real-Time Prediction of Melodies: Continuous Predictability Judgments and Dynamic Models. *Proceedings of the 7th International Conference on Music Perception and Cognition*, (pp. 473–476).

Eerola, T., Louhivuori, J., & Lebaka, E. (2009). Expectancy in Sami Yoiks revisited: The role of data-driven and schema-driven knowledge in the formation of melodic expectations. *Musicae Scientiae*, *13*, 231–272.

Fine, A. B., & Jaeger, T. F. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science*, *37*, 578–591.

Fiser, J., & Aslin, R. N. (2016). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*, 499–504.

Fogel, A. R., Rosenberg, J. C., Lehman, F. M., Kuperberg, G. R., & Patel, A. D. (2015). Studying musical and linguistic prediction in comparable ways: The melodic cloze probability method. *Frontiers in Psychology*, *6*, 247–14.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*, 293–301.

Froyen, V., Feldman, J., & Singh, M. (2015). Bayesian hierarchical grouping: Perceptual grouping as mixture estimation. *Psychological Review*, *122*, 575–597.

Goodman, N. (1976). *Languages of Art*. Indianapolis: Hackett.

Handel, S. (1993). Listening: An introduction to the perception of auditory events. The MIT Press.

Hansen, N. C., & Pearce, M. T. (2014). Predictive uncertainty in auditory sequence processing. *Frontiers in Psychology*, *5*, 1–17.

Hsu, Y. F., Le Bars, S., Hamalainen, J. A., & Waszak, F. (2015). Distinctive representation of mispredicted and unpredicted prediction errors in human electroencephalography. *Journal of Neuroscience*, *35*, 14653–14660.

Huron, D. (2001). Tone and Voice: A Derivation of the Rules of Voice-Leading from Perceptual Principles. *Music Perception: An Interdisciplinary Journal*, *19*, 1–64.

Huron, D. B. (2006). *Sweet Anticipation*. Music and the Psychology of Expectation. MIT Press.

Jackendoff, R. (1992). *Languages of the Mind*. Essays on Mental Representation. Cambridge, MA: MIT Press.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *PNAS*, *105*, 10687–10692.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS ONE*, *7*, e36399–8.

Kim, J. N., Large, E. W., Gwon, Y., & Ashley, R. (2018). The Online Processing of Implied Harmony in the Perception of Tonal Melodies. *Music Perception: An Interdisciplinary Journal*, *35*, 594–606.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, *83*, B35–B42.

Koelsch, S., Schroger, E., & Gunter, T. (2002). Music matters: Preattentive musicality of the human brain. *Psychophysiology*, *39*, 38–48.

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

Krumhansl, C., Toivanen, P., Eerola, T., Toiviainen, P., Jarvinen, T., & Louhivuori, J. (2000). Cross-cultural music cognition: cognitive methodology applied to North Sami yoiks. *Cognition*, *76*, 13–58.

Krumhansl, C. L., Louhivuori, J., Toiviainen, P., Jarvinen, T., & Eerola, T. (1999). Melodic Expectation in Finnish Spiritual Folk Hymns: Convergence of Statistical, Behavioral, and Computational Approaches. *Music Perception: An Interdisciplinary Journal*, *17*, 151–195.

Kuperberg, G. R., & Jaeger, T. F. (2015). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*, 32–59.

Larson, S. (2004). Musical Forces and Melodic Expectations: Comparing Computer Models and Experimental Results. *Music Perception: An Interdisciplinary Journal*, *21*, 457–498.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

Loui, P., & Wessel, D. (2007). Harmonic expectation and affect in Western music: Effects of attention and training. *Perception & Psychophysics*, *69*, 1084–1092.

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

Margulis, E. H. (2005). A model of melodic expectation. *Music Perception: An Interdisciplinary Journal*, *22*, 663–714.

Meyer, L. B. (1956). *Emotion and Meaning in Music*. University of Chicago Press.

Moldwin, T., Schwartz, O., & Sussman, E. S. (2017). Statistical learning of melodic patterns influences the brain's response to wrong notes. *Journal of Cognitive Neuroscience*, *29*, 2114–2122.

Morgan, E. (2016). *Generative and Item-Specific Knowledge of Language*. Ph.D. thesis University of California, San Diego.

Morgan, E., & Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, *157*, 384–402.

Narmour, E. (1989). The "Genetic Code" of Melody: Cognitive Structures Generated by the Implication-Realization Model. *Contemporary Music Review*, *4*, 45–64.

Narmour, E. (1990). *The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model*. Chicago: University of Chicago Press.

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, *6*, 674–681.

Patel, A. D., & Morgan, E. (2016). Exploring Cognitive Relations Between Prediction in Language and Music. *Cognitive Science*, *41*, 303–320.

Pearce, M. T. (2005). *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. Ph.D. thesis City University, London.

Pearce, M. T., Ruiz, M. H., Kapasi, S., Wiggins, G. A., & Bhattacharya, J. (2010). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage*, *50*, 302–313.

Pearce, M. T., & Wiggins, G. A. (2006). Expectation in melody: The influence of context and learning. *Music Perception: An Interdisciplinary Journal*, *23*, 377–405.

Pinker, S. (2000). *Words and Rules*. The Ingredients of Language. New York: Harper Perennial.

Povel, D.-J., & Essens, P. (1985). Perception of Temporal Patterns. *Music Perception: An Interdisciplinary Journal*, *2*, 411–440.

Povel, D.-J., & Jansen, E. (2002). Harmonic factors in the perception of tonal melodies. *Music Perception: An Interdisciplinary Journal*, *20*, 51–85.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: https://www.R-project.org/.

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, *3*, 504–509.

Rock, I., & Palmer, S. (1990). The Legacy of Gestalt Psychology. *Scientific American*, *263*, 84–91.

Rohrmeier, M. (2011). Towards a generative syntax of tonal harmony. *Journal of Mathematics and Music*, *5*, 35–53.

Rohrmeier, M. (2016). Musical Expectancy. *Zeitschrift der Gesellschaft fur Musiktheorie*, *10*, 343–371.

Rohrmeier, M. A., & Koelsch, S. (2012). Predictive information processing in music cognition. A critical review. *International Journal of Psychophysiology*, *83*, 164–175.

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

Sapp, C. (2018). *Songs and Ballads from Nova Scotia*. URL: http://kern.ccarh.org/browse?l=folk/novascotia (accessed January 24, 2018).

Schaffrath, H., & Huron, D. (1995). *The Essen Folksong Collection in Kern Format*. URL: http://kern.humdrum.org/cgi-bin/browse?l=/essen (accessed January 24, 2018).

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

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

Schubotz, R. (2007). Prediction of external events with our motor system: Towards a new framework. *Trends in Cognitive Sciences*, *11*, 211–218.

Sears, D. R. W., Pearce, M. T., Caplin, W. E., & McAdams, S. (2018). Simulating melodic and harmonic expectations for tonal cadences using probabilistic models. *Journal of New Music Research*, *47*, 29–52.

Smith, N., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, (pp. 1637–1642).

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, *82*, 1–17.

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

Temperley, D. (2014). Probabilistic Models of Melodic Interval. *Music Perception: An Interdisciplinary Journal*, *32*, 85–99.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309–318.

Tillmann, B. (2012). Music and Language Perception: Expectations, Structural Integration, and Cognitive Sequencing. *Topics in Cognitive Science*, (pp. 1–17).

Trainor, L. J. (2015). The origins of music in auditory scene analysis and the roles of evolution and culture in musical creation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*, 1–14.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 443–467.

Van Petten, C., & Luka, B. J. (2011). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*, 176–190.

Von Hippel, P., & Huron, D. (2000). Why Do Skips Precede Reversals? The Effect of Tessitura on Melodic Structure. *Music Perception: An Interdisciplinary Journal*, *18*, 59–85.

Vuust, P., Ostergaard, L., Pallesen, K. J., Bailey, C., & Roepstorff, A. (2009). Predictive coding of music - Brain responses to rhythmic incongruity. *CORTEX*, *45*, 80–92.

van der Weij, B., Pearce, M. T., & Honing, H. (2017). A probabilistic model of meter perception: Simulating enculturation. *Frontiers in Psychology*, *8*, 217–18.

Winner, E. (2018). *How Art Works*. A Psychological Exploration. New York: Oxford Univ. Press.

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attn Percept Psychophys*, *79*, 2064–2072.

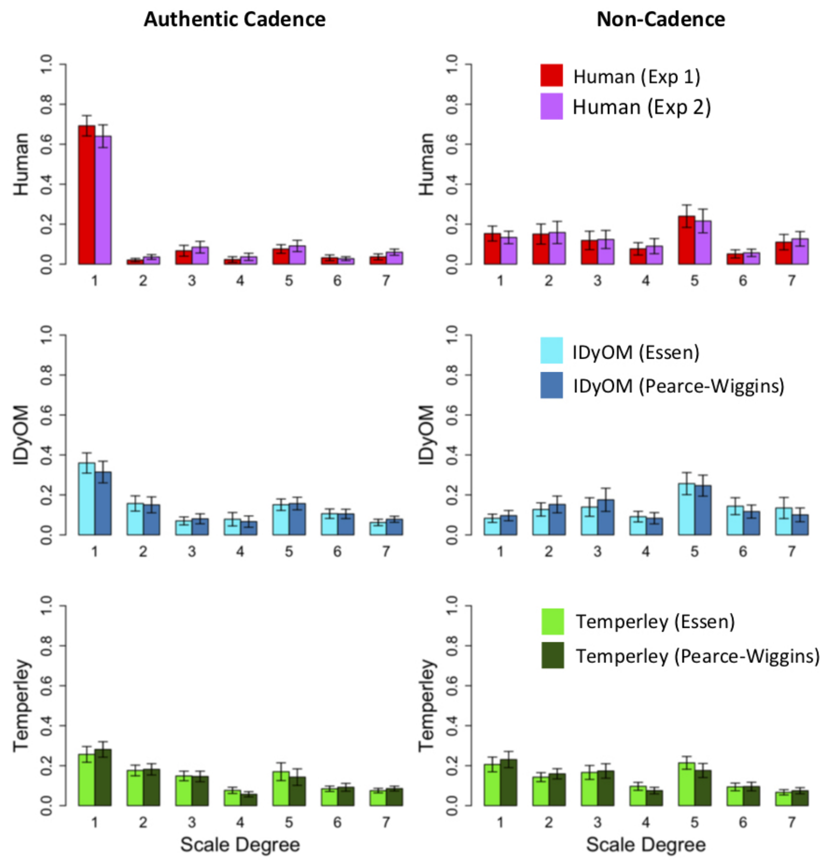

Figure 3: Human data from the melodic cloze experiments (top row) and raw predictions from both models, based on both training sets. y-axes show proportion of responses as given by humans or predicted by models. Error bars show ±2 standard errors (computed over melodies).

Figure 4: Melody NC43 likely causes "stream segregation" wherein listeners perceive two melodic lines.

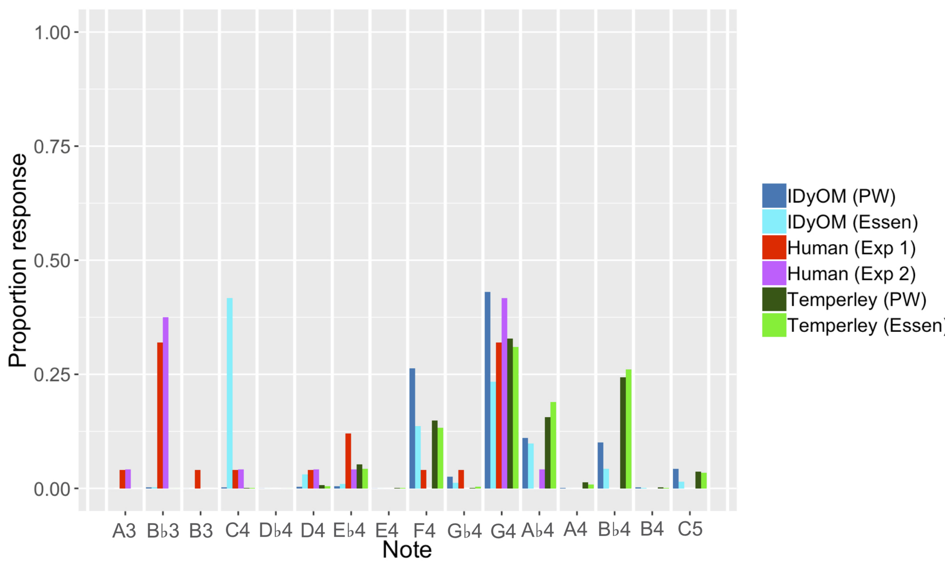

Figure 5: Proportion of human responses and model predictions for melody NC43 (shown in Figure 4), for models trained on the Essen and Pearce-Wiggins (PW) corpora. A substantial proportion of human participants continued this melody with B♭3, which is unpredicted by any of the models.