UNIVERSITY OF CALIFORNIA PRESS
JOURNALS + DIGITAL PUBLISHING

EXPECTATION IN MELODY: THE INFLUENCE OF CONTEXT AND LEARNING

—

MARCUS T. PEARCE & GERAINT A. WIGGINS
Centre for Cognition, Computation and Culture
Goldsmiths College, University of London

THE IMPLICATION-REALIZATION (IR) theory (Narmour, 1990) posits two cognitive systems involved in the generation of melodic expectations: The first consists of a limited number of symbolic rules that are held to be innate and universal; the second reflects the top-down influences of acquired stylistic knowledge. Aspects of both systems have been implemented as quantitative models in research which has yielded empirical support for both components of the theory (Cuddy & Lunny, 1995; Krumhansl, 1995a, 1995b; Schellenberg, 1996, 1997). However, there is also evidence that the implemented bottom-up rules constitute too inflexible a model to account for the influence of the musical experience of the listener and the melodic context in which expectations are elicited. A theory is presented, according to which both bottom-up and top-down descriptions of observed patterns of melodic expectation may be accounted for in terms of the induction of statistical regularities in existing musical repertoires. A computational model that embodies this theory is developed and used to reanalyze existing experimental data on melodic expectancy. The results of three experiments with increasingly complex melodic stimuli demonstrate that this model is capable of accounting for listeners' expectations as well as or better than the two-factor model of Schellenberg (1997).

Received January 12, 2005, accepted November 21, 2005

———

The generation of expectations is recognized as being an especially important factor in music cognition. From a music-analytic perspective, it has been argued that the generation and subsequent confirmation or violation of expectations is critical to aesthetic experience, and the communication of emotion and meaning in music (Meyer, 1956; Narmour, 1990). From a psychological perspective, expectancy has been found to influence recognition memory for music (Schmuckler, 1997), the production of music (Carlsen, 1981; Schmuckler, 1989, 1990; Thompson, Cuddy, & Plaus, 1997; Unyk & Carlsen, 1987), the perception of music (Cuddy & Lunny, 1995; Krumhansl, 1995b; Schellenberg, 1996; Schmuckler, 1989), and the transcription of music (Unyk & Carlsen, 1987). While most empirical research has examined the influence of melodic structure, expectancy in music also reflects the influence of rhythmic and metric structure (Jones, 1987; Jones & Boltz, 1989) as well as harmonic structure (Bharucha, 1987; Schmuckler, 1989).

The present research examines the cognitive mechanisms underlying the generation of melodic expectations. Narmour (1990, 1992) has proposed a detailed and influential theory of expectancy in melody which attempts to characterize the set of implied continuations to an incomplete melodic sequence. According to the theory, the expectations of a listener are influenced by two distinct cognitive systems: first, a bottom-up system consisting of Gestalt-like principles that are held to be innate and universal; and second, a top-down system consisting of style-specific influences on expectancy which are acquired through extensive exposure to music in a given style. Krumhansl (1995b) has formulated the bottom-up system of the IR theory as a quantitative model, consisting of a small set of symbolic rules. This model has formed the basis of a series of empirical studies, which have examined the degree to which the expectations of listeners conform to the predictions of the IR theory and have led to several different formulations of the principles comprising the bottom-up component of the model.

While this body of research suggests that the expectations of listeners in a given experiment may be accounted for by some collection of principles intended to reflect the bottom-up and top-down components of Narmour's theory, the present research is motivated by empirical data that question the existence of a small set of universal bottom-up rules that determine, in part, the expectations of a listener. According to the theory presented here, expectancy in melody can be accounted for entirely in terms of the induction of statistical regularities in sequential melodic structure without recourse to an independent system of innate symbolic predisposi-

tions. While innate constraints on music perception certainly exist, it is argued that they are unlikely to be found in the form of rules governing sequential dependencies between musical events. According to the account developed here, patterns of expectation that do not vary between musical styles are accounted for in terms of simple regularities in music whose ubiquity may be related to the constraints of physical performance. If this is the case, there is no need to make additional (and problematic) assumptions about innate representations of sequential dependencies between perceived events (Elman et al., 1996).

The specific goals of this research are twofold. The first is to examine whether models of melodic expectancy based on statistical learning are capable of accounting for the patterns of expectation observed in empirical behavioral research. If such models can account for the behavioral data as well as existing implementations of the IR theory, there would be no need to invoke symbolic rules as universal properties of the human cognitive system. To the extent that such models can be found to provide a more powerful account of the behavioral data, the IR theory (as currently implemented) may be viewed as an inadequate cognitive model of melodic expectancy by comparison. Instead of representing innate and universal constraints of the perceptual system, the bottom-up principles may be taken to represent a formalized approximate description of the mature behavior of a cognitive system of inductive learning. The second goal of the present research is to undertake a preliminary examination of the kinds of melodic feature that afford regularities capable of supporting the acquisition of the patterns of expectation exhibited by listeners.

In order to achieve these goals, a computational model embodying the proposed theory of expectancy is developed and used to predict empirical data on the patterns of melodic expectation exhibited by listeners. The fit of the model to the behavioral data is compared to that obtained with a quantitative formulation of the IR theory consisting of two bottom-up principles (Schellenberg, 1997).

The question of distinguishing acquired and inherited components of behavior is a thorny one, all the more so in relation to the perception of cultural artifacts (which are both created and appreciated through the application of the human cognitive system). Following Cutting, Bruno, Brady, and Moore (1992), three criteria are used to compare the two cognitive models of melodic expectation. The first criterion is *scope*, which refers to the degree to which a theory accounts for a broad range of experimental data elicited in a variety of contexts. In order to evaluate the scope of the two models, the extent to which they account for the patterns of expectation exhibited by listeners is examined and compared in three experiments which investigate expectations elicited in the context of increasingly complex melodic stimuli. Each experiment also incorporates analyses of more detailed hypotheses concerning the melodic features that afford regularities capable of supporting the acquisition of the observed patterns of expectation.

The second criterion introduced by Cutting et al. (1992) is *selectivity*, which refers to the degree to which a theory accounts specifically for the data of interest and does not predict unrelated phenomena. In order to compare the models on the basis of selectivity, the ability of each model to account for random patterns of expectation is assessed in each experiment.

The third criterion discussed by Cutting et al. (1992) is the *principle of parsimony* (or *simplicity*): a general methodological heuristic expressing a preference for the more parsimonious of two theories that each account equally well for observed data. Although the precise operational definition of parsimony is a point of debate in the philosophy of science, variants of the heuristic are commonly used in actual scientific practice (Nolan, 1997; Popper, 1959; Sober, 1981). This provides some evidence that the principle is normative, that is, that it actually results in successful theories. Further evidence along these lines is provided by the fact that simplicity is commonly used a heuristic bias in machine learning (Mitchell, 1997) and for hypothesis selection in abductive reasoning (Paul, 1993).

Furthermore, quantifying the principle of parsimony in terms of algorithmic information theory demonstrates that simple encodings of a set of data also provide the most probable explanations for that data (Chater, 1996, 1999; Chater & Vitányi, 2003). In the closely related field of Bayesian inference, it is common to compare models according to their simplicity, measured as a function of the number of free parameters they possess and the extent to which these parameters need to be finely tuned to fit the data (Jaynes, 2003; MacKay, 2003). Chater (1999) presents simplicity as a rational analysis of perceptual organization on the basis of these normative justifications together with evidence that simple representations of experience are preferred in perception and cognition. Although this application of simplicity is not a primary concern in the present research, we touch on it again as a justification for preferring small feature sets and when discussing the results of Experiment 3.

In psychology (as in many other scientific fields), the relative parsimony of comparable models is most commonly defined in terms of the number of free

parameters in each model (Cutting et al., 1992). Here, however, we use the principle in a more general sense where the existence of a theoretical component assumed by one theory is denied leading to a simpler theory (Sober, 1981). To the extent that the theory of inductive learning is comparable to the top-down component of the IR theory (and in the absence of specific biological evidence for the innateness of the bottom-up principles), the former theory constitutes a more parsimonious description of the cognitive system than the latter since additional bottom-up constraints *assumed* to constitute part of the cognitive system are replaced by equivalent constraints *known* to exist in the environment. In order to test this theoretical position, we examine the extent to which the statistical model subsumes the function of the two-factor model of expectancy in accounting for the behavioral data in each experiment.

Finally, the article concludes with a general discussion of the experimental results, their implications, and some promising directions for further development of the theory.

Background

The Implication-Realization Theory

Building on the work of Meyer (1956, 1973), Narmour (1990, 1991, 1992) has developed a complex theory of melody perception called the *Implication-Realization* (IR) theory. The theory posits two distinct perceptual systems—the *bottom-up* and *top-down systems* of melodic implication. While the principles of the former are held to be hardwired, innate, and universal, the principles of the latter are held to be learned and hence dependent on musical experience.

> The top-down system is flexible, variable and empirically driven. . . . In contrast, the bottom-up mode constitutes an automatic, unconscious, preprogrammed, "brute" system. (Narmour, 1991, p. 3)

In the bottom-up system, the rhythmic, metric, tonal, and intervallic properties of a sequence of melodic intervals determine the degree of *closure* conveyed by the sequence. While strong closure signifies the termination of ongoing melodic structure, an unclosed or *implicative* interval generates expectations for the following interval, which is termed the *realized interval*. The expectations generated by implicative intervals are described by Narmour (1990) in terms of several principles of implication which are influenced by the Gestalt principles of proximity, similarity, and good

continuation. In particular, according to the theory, small melodic intervals imply a *process* (the realized interval is in the same direction as the implicative interval and will be similar in size) while large melodic intervals imply a *reversal* (the realized interval is in a different direction to the implicative interval and is smaller in size).

Although the theory is presented in a highly analytic manner, it has psychological relevance because it advances hypotheses about general perceptual principles that are precisely and quantitatively specified and therefore amenable to empirical investigation (Krumhansl, 1995b; Schellenberg, 1996). In particular, a number of different authors have expressed the bottom-up system as a quantitative model consisting of a number of symbolic principles. The following description of the principles of the bottom-up system is based on an influential summary by Krumhansl (1995b). Some of these principles operate differently for small and large intervals which are defined to be those of five semitones or less and seven semitones or more respectively. The tritone is considered by Narmour (1990) to be a threshold interval assuming the function of a small or large interval (i.e., implying continuation or reversal) depending on the context.

Registral direction states that small intervals imply continuations in the same registral direction whereas large intervals imply a change in registral direction. The application of the principle to small intervals is related to the Gestalt principle of good continuation.

Intervallic difference states that small intervals imply a subsequent interval that is similar in size (± 2 semitones if registral direction changes and ± 3 semitones if direction continues), while large intervals imply a consequent interval that is smaller in size (at least three semitones smaller if registral direction changes and at least four semitones smaller if direction continues). This principle can be taken as an application of the Gestalt principles of similarity and proximity for small and large intervals respectively.

Registral return is a general implication for a return to the pitch region (± 2 semitones) of the first tone of an implicative interval in cases where the realized interval reverses the registral direction of the implicative interval. Krumhansl (1995b) coded this principle as a dichotomy although Narmour (1990) distinguishes between *exact* and *near* registral return suggesting that the principle be graded as a function of the size of the interval between the realized tone and the first tone of the implicative interval (Schellenberg, 1996; Schellenberg, Adachi, Purdy, & McKinnon, 2002). This principle can be viewed as an

application of the Gestalt principles of proximity in terms of pitch and similarity in terms of pitch interval.

Proximity describes a general implication for small intervals (five semitones or less) between any two tones. The implication is graded according to the absolute size of the interval. This principle can be viewed as an application of the Gestalt principle of proximity.

Closure is determined by two conditions: first, a change in registral direction; and second, movement to a smaller-sized interval. Degrees of closure exist corresponding to the satisfaction of both, one or neither of the conditions.

In this encoding, the first three principles (registral direction, intervallic difference, and registral return) assume dichotomous values while the final two (proximity and closure) are graded (Krumhansl, 1995b). Although the bottom-up IR principles are related to generic Gestalt principles, they are parametrized and quantified in a manner specific to music.

Narmour (1990) makes explicit use of the principles of registral direction and intervallic difference to derive a complete set of 12 basic melodic structures each consisting of an implicative and a realized interval. These basic structures are differentiated by the size and direction of the realized interval relative to those of the implicative interval and the absolute size of the implicative interval. In an experimental study of the IR theory, Krumhansl (1995b) reports only limited support for the basic melodic structures suggesting that expectations depend not only on registral direction and intervallic difference but also on the principles of proximity, registral return, and closure, which are less explicitly formulated in the original presentation of the IR theory (Krumhansl, 1995b).

In other respects, the quantitatively formulated model developed by Krumhansl (1995b) lacks some of the more complex components of the IR theory. For example, Narmour (1992) presents a detailed analysis of how the basic melodic structures combine together to form longer and more complex structural patterns of melodic implication within the IR theory. Furthermore, tones emphasized by strong closure are transformed to a higher level of structural representation which may retain some of the registral implications of the lower level. Krumhansl (1997) has found some empirical support for the psychological validity of higher-level implications in experiments with specially constructed melodic sequences. Finally, although quantitative implementations have tended to focus on the *parametric scales* of registral direction and interval size, the IR theory also includes detailed treatment of other parametric

scales such as duration, metric emphasis, and harmony (Narmour, 1990, 1992).

The IR theory also stresses the importance of top-down influences on melodic expectancy. The top-down system is acquired on the basis of musical experience and, as a consequence, varies across musical cultures and traditions. The influences exerted by the top-down system include both *extraopus* knowledge about style-specific norms such as diatonic interpretations, tonal and metrical hierarchies, and harmonic progressions and *intraopus* knowledge about aspects of a particular composition such as distinctive motivic and rhythmic patterns. Bharucha (1987) makes a similar distinction between *schematic* and *veridical* influences on expectancy: While the former are influenced by schematic representations of typical musical relationships acquired through extensive exposure to a style, the latter are aroused by the activation of memory traces for specific pieces or prior knowledge of what is to come. Finally, the top-down system may generate implications that conflict with and potentially override those generated by the bottom-up system. Efforts to develop quantitative implementations of the IR theory have tended to focus on the bottom-up system with the top-down system represented only by relatively simple quantitative predictors.

It is important to emphasize that the present research is primarily concerned with those concrete implementations of the IR theory that, although they lack much of the music-analytic detail of Narmour's theory, have been examined in an empirical, psychological context. Although Narmour considered the five principles summarized above to be "a fair representation of his model" (Schellenberg, 1996, p. 77) and refers the reader to Krumhansl (1995b) among others for "evaluations of the model" (Narmour, 1999, p. 446), the present research is relevant to the IR theory of Narmour (1990, 1992) only to the extent that the concrete implementations examined are viewed as representative of the basic tenets of the theory. The IR theory has been the subject of several detailed reviews published in the psychological and musicological literature (Cross, 1995; Krumhansl, 1995b; Thompson, 1996) to which the reader is referred for more thorough summaries of its principal features.

Empirical Studies of Melodic Expectancy

Overview

Expectancy in music has been studied in experimental settings from a number of perspectives including the influence of rhythmic (Jones, 1987; Jones & Boltz, 1989), melodic (Cuddy & Lunny, 1995; Krumhansl, 1995b)

and harmonic structure (Bharucha, 1987; Schmuckler, 1989). A variety of experimental paradigms have been employed to study expectancy including rating completions of musical contexts (Cuddy & Lunny, 1995; Krumhansl, 1995a; Schellenberg, 1996), generating continuations to musical contexts (Carlsen, 1981; Schmuckler, 1989; Thompson et al., 1997; Unyk & Carlsen, 1987), classifying and remembering musical fragments (Schmuckler, 1997), reaction time experiments (Aarden, 2003; Bharucha & Stoeckig, 1986), and continuous response methodologies (Eerola, Toiviainen, & Krumhansl, 2002). Although expectancy in music has been shown to operate in a number of different contexts over a number of different parameters and structural levels in music, this review is restricted to studies of expectancy in melodic music and, in particular, those which have specifically addressed the claims of the IR theory. The following two sections present reviews of empirical research examining the predictions of the bottom-up and top-down components of the theory.

The Bottom-up System

Cuddy and Lunny (1995) tested the bottom-up principles of the IR theory (as quantified by Krumhansl, 1995b) against goodness-of-fit ratings collected for continuation tones following a restricted set of two-tone melodic beginnings (see also Experiment 1). A series of multiple regression analyses supported the inclusion of intervallic difference, proximity, and registral return in a theory of melodic expectancy. Support was also found for a revised version of registral direction, which pertains to large intervals only, and an additional bottom-up principle of pitch height, based on the observation that ratings tended to increase as the pitch height of the continuation tone increased. No support was found for the bottom-up principle of closure.

Krumhansl (1995a) repeated the study of Cuddy and Lunny (1995) with 16 musically trained American participants using a more complete set of two-tone contexts ranging from a descending major seventh to an ascending major seventh. Analysis of the results yielded support for modified versions of proximity, registral return, and registral direction but not closure or intervallic difference. In particular, the results supported a modification of proximity such that it is linearly graded over the entire range of intervals used and a modification of registral return such that it varies as a linear function of the proximity of the third tone to the first. Finally, the principle of registral direction was supported by the analysis except for the data for the major seventh which carried strong implications for

octave completion (see also Carlsen, 1981). Support was also found for two extra principles that distinguish realized intervals forming octaves and unisons respectively. Krumhansl (1995a) also examined the effects of bottom-up psychophysical principles finding support for predictors coding the consonance of a tone with the first and second tones of the preceding interval (based on empirical and theoretical considerations).

Other experimental studies have extended these findings to expectations generated by exposure to melodic contexts from existing musical repertoires. Krumhansl (1995b) reports a series of three experiments: The first used eight melodic fragments taken from British folk songs, diatonic continuation tones, and 20 American participants of whom 10 were musically trained and 10 untrained (see also Experiment 2); the second used eight extracts from Webern's *Lieder* (Opus 3, 4, and 15), chromatic continuation tones, and 26 American participants generally unfamiliar with the atonal style of whom 13 were musically trained and 13 untrained; and the third used 12 melodic fragments from Chinese folk songs, pentatonic continuation tones, and 16 participants of whom 8 were Chinese and 8 American. All the melodic contexts ended on an implicative interval and all continuation tones were within a two-octave range centered on the final tone of the context. Analysis of the results yielded support for all of the bottom-up principles (with the exception of intervallic difference for the second experiment). Overall, the weakest contribution was made by intervallic difference and the strongest by proximity. With the exception of the first experiment, support was also found for the unison principle of Krumhansl (1995a).

Schellenberg (1996) argued that the bottom-up models discussed above are overspecified and contain redundancy due to collinearities between their component principles. As a result, the theory may be expressed more simply and parsimoniously without loss of predictive power. Support was found for this argument in an independent analysis of the experimental data reported by Krumhansl (1995b) using a model consisting of registral return, registral direction revised such that it applies only to large intervals (although quantified in a different manner to the revision made by Cuddy & Lunny, 1995), and a revised version of proximity (similar in spirit, though quantitatively different, to the revision made by Krumhansl, 1995a). In a further experiment, Schellenberg (1997) applied principal components analysis to this revised model with the resulting development of a two-factor model. The first factor is the principle of proximity as revised by Schellenberg (1996); the second, *pitch reversal,* is an additive combination of the principles

of registral direction (revised) and registral return. This model is considerably simpler and more parsimonious than Schellenberg's revised model and yet does not compromise the predictive power of that model in accounting for the data obtained by Krumhansl (1995b) and Cuddy and Lunny (1995).

Similar experiments with Finnish spiritual folk hymns (Krumhansl, Louhivuori, Toiviainen, Järvinen, & Eerola, 1999) and indigenous folk melodies (yoiks) of the Sami people of Scandinavia (Krumhansl et al., 2000) have, however, questioned the cross-cultural validity of such revised models. In both studies, it was found that the model developed by Krumhansl (1995a) provided a much better fit to the data than those of Krumhansl (1995b) and Schellenberg (1996, 1997). By contrast, Schellenberg et al. (2002) have found the opposite to be true in experiments with adults and infants in a task involving the rating of continuation tones following contexts taken from Acadian (French Canadian) folk songs. They suggest that the difference may be attributable partly to the fact that none of the musical contexts used in the experiments of Krumhansl et al. (1999, 2000) ended in unambiguously large and implicative intervals (Schellenberg et al., 2002, p. 530). While Schellenberg et al. (2002) and Krumhansl et al. (1999) found strong support for the principle of proximity with only limited influence of registral return and intervallic difference, Krumhansl et al. (2000) found the strongest bottom-up influence came from the principle of intervallic difference with weak support for the principles of proximity and registral return. The consonance predictors of Krumhansl (1995a) made a strong contribution to both models especially in the case of the folk hymns (Krumhansl et al., 1999, 2000).

According to the IR theory, the principles of the bottom-up system exert a consistent influence on expectations regardless of the musical experience of the listener and the stylistic context notwithstanding the fact that the expectations actually generated are predicted to be subject to these top-down influences. Indirect support for this claim comes in the form of high correlations between the responses of musically trained and untrained participants (Cuddy & Lunny, 1995; Schellenberg, 1996) and between the responses of groups with different degrees of familiarity with the musical style (Eerola, 2004a; Krumhansl et al., 1999, 2000; Schellenberg, 1996). Regardless of the cognitive mechanisms underlying the generation of melodic expectations, it is clear that they tend to exhibit a high degree of similarity across levels of music training and familiarity. More direct evidence is provided by qualitatively similar degrees of influence of the bottom-up principles on the expectations of

musically trained and untrained participants (Cuddy & Lunny, 1995; Schellenberg, 1996) and across levels of relevant stylistic experience (Krumhansl et al., 1999; Schellenberg, 1996). These findings have typically been interpreted as support for the universality of the bottom-up principles.

However, there are several reasons to question this conclusion. First, other research on melodic expectancy has uncovered differences across levels of training. von Hippel (2002), for example, conducted an experiment in which trained and untrained participants were asked to make prospective contour judgments for a set of artificially generated melodies. While the expectations of the trained listeners exhibited the influence of pitch reversal and *step momentum* (the expectation that a melody will maintain its registral direction after small intervals), the responses of the untrained listeners exhibited significantly weaker influences of these principles. Furthermore, in a study of goodness-of-fit ratings of single intervals as melodic openings and closures, Vos and Pasveer (2002) found that the responses of untrained listeners exhibited a greater influence of intervallic direction than those of the trained listeners.

Second, it must be noted that the empirical data cover a limited set of cultural groups and that differences in observed patterns of expectation related to cultural background have been found (Carlsen, 1981). Furthermore, some studies have uncovered cross-cultural differences in the strength of influence of the bottom-up principles on expectancy. Krumhansl et al. (2000), for example, found that the correlations of the predictors for intervallic difference, registral return, and proximity were considerably stronger for the Western listeners than for the Sami and Finnish listeners. Eerola (2004a) made similar observations in a replication of this study with traditional healers from South Africa.

Third, the influence of the bottom-up principles appears to vary with the musical stimuli used. Krumhansl et al. (2000) note that while the Finnish listeners in their study of expectancy in Sami folk songs exhibited a strong influence of consonance, the Finnish listeners in the earlier study of expectancy in Finnish hymns (Krumhansl et al., 1999) exhibited a weaker influence of consonance in spite of having a similar musical background. Krumhansl et al. (2000) suggest that this may indicate that the Finnish listeners in their study adapted their judgments to the relatively large number of consonant intervals present in their experimental materials. More generally, the research reviewed in this section diverges significantly in the support found for the original bottom-up principles, revised

versions of these principles, and new principles. The most salient differences between the studies, and the most obvious causes of such discrepancies, are the musical contexts used to elicit expectations. Krumhansl et al. (2000, p. 41) conclude that "musical styles may share a core of basic principles, but that their relative importance varies across styles."

The influence of melodic context on expectations has been further studied by Eerola et al. (2002) who used a continuous response methodology to collect participants' continuous judgments of the predictability of melodies (folk songs, songs by Charles Ives, and isochronous artificially generated melodies) simultaneously as they listened to them. The predictability ratings were analyzed using three models: first, the IR model; second, a model based on the entropy of a monogram distribution of pitch intervals with an exponential decay within a local sliding window (the initial distribution was derived from an analysis of the Essen Folk Song Collection, Schaffrath, 1992, 1994); and third, a variant of the second model in which the pitch class distribution was used and was initialized using the key profiles of Krumhansl and Kessler (1982). The results demonstrated that the second model and, in particular, the third model accounted for much larger proportions of the variance in the predictability data than the IR model while a linear combination of the second and third models improved the fit even further (Eerola, 2004b). It was argued that the success of these models was a result of their ability to account for the data-driven influences of melodic context.

Finally, it is important to note that universality or ubiquity of patterns of behavior does not imply innateness. To the extent that the bottom-up principles capture universal patterns of behavior, they may reflect the influence of long-term informal exposure to simple and ubiquitous regularities in music (Schellenberg, 1996; Thompson et al., 1997). In accordance with this position, Bergeson (1999) found that while adults are better able to detect a pitch change in a melody that fulfills expectations according to the IR theory (Narmour, 1990) than in one that does not, 6- and 7-month-old infants do not exhibit this difference in performance across conditions. In addition, Schellenberg et al. (2002) report experiments examining melodic expectancy in adults and infants (covering a range of ages) using experimental tasks involving both rating and singing continuation tones to supplied melodic contexts. The data were analyzed in the context of the IR model as originally formulated by Schellenberg (1996) and the two-factor model of Schellenberg (1997). The results demonstrate that expectations were better explained by both models with increasing age and musical exposure. While consecutive pitch proximity (Schellenberg, 1997) was a strong influence for all listeners, the influence of more complex predictors such as pitch reversal (Schellenberg, 1997) and registral return (Schellenberg, 1996) only became apparent with the older listeners. Schellenberg et al. (2002) conclude with a discussion of possible explanations for the observed developmental changes in melodic expectancy: First, they may reflect differences between infant-directed speech and adult-directed speech; second, they may reflect general developmental progressions in perception and cognition (e.g., perceptual differentiation and working or sensory memory), which exert influence across domains and modalities; and third, they may reflect increasing exposure to music and progressive induction of increasingly complex regularities in that music.

The Top-down System

In addition to studying the bottom-up principles of the IR theory, research has also examined some putative top-down influences on melodic expectation many of which are based on the key profiles of perceived tonal stability empirically quantified by Krumhansl and Kessler (1982). Schellenberg (1996) and Krumhansl (1995b), for example, found support for the inclusion in a theory of expectancy of a tonality predictor based on the key profile for the major or minor key of the melodic fragment. Cuddy and Lunny (1995) examined the effects of several top-down tonality predictors. The first consisted of four tonal hierarchy predictors similar to those of Schellenberg (1996) and Krumhansl (1995b) based on the major and minor key profiles for the first and second tones of the context interval. The second, *tonal strength*, was based on the assumption that the rating of a continuation tone would be influenced by the degree to which the pattern of three tones suggested a tonality. The key-finding algorithm developed by Krumhansl and Schmuckler (Krumhansl, 1990) was used to rate each of the patterns for tonal strength. The third tonality predictor, *tonal region*, was derived by listing all possible major and minor keys in which each implicative interval was diatonic and coding each continuation tone according to whether it represented a tonic of one of these keys. Support was found for all of these top-down influences although it was also found that the predictors for tonal hierarchy could be replaced by tonal strength and tonal region without loss of predictive power. Krumhansl (1995a) extended the tonal region predictor developed by Cuddy and Lunny (1995) by averaging the key profile data for all keys in which

the two context tones are diatonic. Strong support was found for the resulting predictor variable for all context intervals except for the two (ascending and descending) tritones. In contrast, no support was found for the tonal strength predictor of Cuddy and Lunny (1995).

While neither Cuddy and Lunny (1995) nor Schellenberg (1996) found any effect of music training on the influence of top-down tonality predictors, Vos and Pasveer (2002) found that the consonance of an interval (based on music-theoretical considerations) influenced the goodness-of-fit judgments of the trained listeners to a much greater extent than those of the untrained listeners in their study of intervals as candidates for melodic openings and closures.

In a further analysis of their data, Krumhansl et al. (1999) sought to distinguish between schematic and veridical top-down influences on expectations (Bharucha, 1987). The schematic predictors were the two-tone continuation ratings obtained by Krumhansl (1995a) and the major and minor key profiles (Krumhansl & Kessler, 1982). The veridical predictors consisted of monogram, digram, and trigram distributions of tones in the entire corpus of spiritual folk hymns and a predictor based on the correct continuation tone. It was found that the schematic predictors showed significantly stronger effects for the nonexperts in the study than the experts. In contrast, veridical predictors such as monogram and trigram distributions and the correct next tone showed significantly stronger effects for the experts than for the nonexperts. Krumhansl et al. (2000) found similar effects in their study of North Sami yoiks and showed that these effects were related to familiarity with individual pieces used in the experiment. These findings suggest that increasing familiarity with a given stylistic tradition tends to weaken the relative influence of top-down schematic knowledge of Western tonal-harmonic music on expectancy and increase the relative influence of specific veridical knowledge of the style.

There is some evidence, however, that the rating of continuation tones may elicit schematic tonal expectations specifically related to melodic closure since the melody is paused to allow the listener to respond. Aarden (2003) reports an experiment in which participants were asked to make retrospective contour judgments for each event in a set of European folk melodies. Reaction times were measured as an indication of the strength and specificity of expectations under the hypothesis that strong and accurate expectations facilitate faster responses (see also Bharucha & Stoeckig, 1986). The resulting data were analyzed using the two-factor model of Schellenberg (1997). While a tonality predictor based on the key profiles of Krumhansl and Kessler

(1982) made no significant contribution to the model, a monogram model of pitch frequency in the Essen Folk Song Collection (Schaffrath, 1992, 1994) did prove to be a significant predictor. In a second experiment, participants were presented with a counter indicating the number of tones remaining in the melody and were asked to respond only to the final tone. In this case, the Krumhansl and Kessler tonality predictor, which bears more resemblance to the distribution of phrase-final tones than that of all melodic tones in the Essen Folk Song Collection, made a significant contribution to the model. On the basis of these results, Aarden (2003) argues that the schematic effects of tonality may be limited to phrase endings whereas data-driven factors, directly reflecting the structure and distribution of tones in the music, have more influence in melodic contexts that do not imply closure.

Finally, it is worth noting that the top-down tonality predictors that have been examined in the context of modeling expectation have typically been rather simple. In this regard, Povel and Jansen (2002) report experimental evidence that goodness ratings of entire melodies depend not so much on the overall stability of the component tones (Krumhansl & Kessler, 1982) but the ease with which the listener is able to form a harmonic interpretation of the melody in terms of both the global harmonic context (key and mode) and the local movement of harmonic regions. The latter process is compromised by the presence of nonchord tones to the extent that they cannot be assimilated by means of anchoring (Bharucha, 1984) or by being conceived as part of a run of melodic steps. Povel and Jansen (2002) argue that the harmonic function of a region determines the stability of tones within that region and sets up expectations for the resolution of unstable tones.

Summary

While the results of many of the individual studies reviewed in the foregoing sections have been interpreted in favor of the IR theory, the overall pattern emerging from this body of research suggests some important qualifications to this interpretation. Empirical research has demonstrated that some collection of quantitatively formulated principles based on the bottom-up IR system can generally account rather well for the patterns of expectation observed in a given experiment but it is also apparent that any such set constitutes too inflexible a model to fully account for the effects of differences across experimental settings in terms of the musical experience of the listeners and the melodic contexts in which expectations are elicited. Regarding the top-down

system, empirical research suggests that the expectations of listeners show strong effects of schematic factors such as tonality although the predictors typically used to model these effects may be too simple and inflexible to account for the effects of varying the context in which expectations are elicited.

Statistical Learning of Melodic Expectancy

The Theory

A theory of the cognitive mechanisms underlying the generation of melodic expectations is presented here. It is argued that this theory is capable of accounting more parsimoniously for the behavioral data than the quantitative formulations of the IR theory while making fewer assumptions about the cognitive mechanisms underlying the perception of music. From the current perspective, the quantitatively formulated principles of the IR theory provide a descriptive, but not explanatory, account of expectancy in melody: They describe human behavior at a general level but do not account for the cognitive mechanisms underlying that behavior. To the extent that the two theories produce similar predictions, they are viewed as lying on different levels of explanation (Marr, 1982; McClamrock, 1991). Both bottom-up and top-down components of the quantitatively formulated IR models have been found to provide an inadequate account of the detailed influences of musical experience and musical context on melodic expectancy. The theory proposed here is motivated by the need to formulate a more comprehensive account of these influences.

In particular, the present theory questions the need, and indeed the validity, of positing a distinction between bottom-up and top-down influences on expectation, and especially the claim that the principles of the bottom-up system reflect innately specified representations of sequential dependencies between musical events. According to the theory, the bottom-up principles of the IR theory constitute a description of common regularities in music which are acquired as mature patterns of expectation through extensive exposure to music. Rather than invoking innate representational rules (such as the bottom-up principles and the basic melodic structures of the IR theory), this theory invokes innate general-purpose learning mechanisms which impose architectural rather than representational constraints on cognitive development (Elman et al., 1996). Given exposure to appropriate musical stimuli, these learning mechanisms can acquire domain-specific representations and behavior which is approximated by the prin-

ciples of the IR theory (see also Bharucha, 1987; Gjerdingen, 1999).

It is hypothesized that the bottom-up principles of the quantitatively formulated IR models (as well as other proposed bottom-up influences on expectancy) reflect relatively simple musical regularities which display a degree of pan-stylistic ubiquity. To the extent that this is the case, these bottom-up IR principles are regarded as formalized approximate descriptions of the mature behavior of a cognitive system that acquires representations of the statistical structure of the musical environment. On the other hand, top-down factors, such as tonality, reflect the induction of rather more complex musical structures which show a greater degree of variability between musical styles. If this is indeed the case, a single learning mechanism may be able to account for the descriptive adequacy of some of the bottom-up principles across degrees of expertise and familiarity as well as for differences in the influence of other bottom-up principles and top-down factors. By replacing a small number of symbolic rules with a general-purpose learning mechanism, the theory can account more parsimoniously for both consistent and inconsistent patterns of expectation between groups of listeners on the basis of differences in prior musical exposure, the present musical context, and the relative robustness of musical regularities across stylistic traditions.

Supporting Evidence

We shall discuss existing evidence that supports the theory in terms of two questions: Are the regularities in music sufficient to support the acquisition of the experimentally observed patterns of melodic expectation? And: Is there any evidence that listeners possess cognitive mechanisms capable of acquiring such behavior through exposure to music?

Regarding the first question, research suggests that expectancy operates very similarly in tasks that elicit ratings of continuations to supplied melodic contexts and tasks that elicit spontaneous production of continuations to melodic contexts (Schellenberg, 1996; Schmuckler, 1989, 1990; Thompson et al., 1997). If the perception and production of melodies are influenced by similar principles, it is pertinent to ask whether existing repertoires of compositions also reflect such influences of melodic implication. Thompson and Stainton (1996, 1998) have examined the extent to which the bottom-up principles of the IR theory are satisfied in existing musical repertoires including the soprano and bass voices of chorales harmonized by J. S. Bach,

melodies composed by Schubert, and Bohemian folk melodies. Preliminary analyses indicated that significant proportions of implicative intervals satisfy the principles of intervallic difference, registral return, and proximity while smaller proportions satisfied the other bottom-up principles. The proportions were highly consistent across the three datasets. Furthermore, a model consisting of the five bottom-up principles accounted for much of the variance in the pitch of tones following implicative intervals in the datasets (as well as closural intervals in the Bohemian folk melodies—Thompson & Stainton, 1998). With the exception of intervallic difference for the Schubert dataset, all five principles contributed significantly to the predictive power of the model. These analyses demonstrate that existing corpora of melodic music contain regularities that tend to follow the predictions of the IR theory and that are, in principle, capable of supporting the acquisition of patterns of expectation that accord with its principles.

Given these findings, an argument can be made that the observed regularities in music embodied by the bottom-up IR principles reflect universal physical constraints of performance rather than attempts to satisfy universal properties of the perceptual system. Examples of such constraints include the relative difficulty of singing large intervals accurately and the fact that large intervals will tend toward the limits of a singer's vocal range (Russo & Cuddy, 1999; Schellenberg, 1997). von Hippel and Huron (2000) report a range of experimental evidence supporting the latter as an explanation of *post-skip reversals* (*cf.* the principles of registral direction and registral return of Krumhansl, 1995b), which they account for in terms of *regression toward the mean* necessitated by tessitura. In one experiment, for example, it was found that evidence for the existence of post-skip reversals in a range of musical styles is limited to those skips (intervals of three semitones or more) that cross or move away from the median pitch of a given corpus of music. When skips approach the median pitch or land on it, there is no significant difference in the proportions of continuations and reversals of registral direction. In spite of this, von Hippel (2002) found that the expectations of listeners actually reflect the influence of perceived post-skip reversals suggesting that patterns of expectation are acquired as heuristics representing simplified forms of more complex regularities in music.

We turn now to the question of whether the cognitive mechanisms exist to acquire the observed patterns of melodic expectation through exposure to existing music. Saffran, Johnson, Aslin, and Newport (1999) have ele-

gantly demonstrated that both adults and 8-month-old infants are capable of learning to segment continuous tone sequences on the basis of differential transitional probability distributions of tones within and between segments. On the basis of these and similar results with syllable sequences, Saffran et al. (1999) argue that infants and adults possess domain general learning mechanisms that readily compute transitional probabilities on exposure to auditory sequences. Furthermore, Oram and Cuddy (1995) conducted a series of experiments in which continuation tones were rated for musical fit in the context of artificially constructed sequences of pure tones in which the tone frequencies were carefully controlled. The continuation tone ratings of both trained and untrained listeners were significantly related to the frequency of occurrence of the continuation tone in the context sequence. Cross-cultural research has also demonstrated the influence of tone distributions on the perception of music (Castellano, Bharucha, & Krumhansl, 1984; Kessler, Hansen, & Shepard, 1984; Krumhansl et al., 1999). In particular, Krumhansl et al. (1999) found significant influences of second order distributions on the expectations of the expert listeners in their study.

There is also evidence that listeners are sensitive to statistical regularities in the size and direction of pitch intervals in the music they are exposed to. In a statistical analysis of a large variety of Western melodic music, for example, Vos and Troost (1989) found that smaller intervals tend to be of a predominantly descending form while larger ones occur mainly in ascending form. A behavioral experiment demonstrated that listeners are able to correctly classify artificially generated patterns that either exhibited or failed to exhibit the regularity. Vos and Troost consider two explanations for this result: first, that it is connected with the possibly universal evocation of musical tension by ascending large intervals and of relaxation by descending small intervals (Meyer, 1973); and second, that it reflects overlearning of conventional musical patterns. Vos and Troost do not strongly favor either account, each of which depends on the experimentally observed sensitivity of listeners to statistical regularities in the size and direction of melodic intervals.

The Model

The theory of melodic expectancy presented above predicts that it should be possible to design a statistical learning algorithm possessing no prior knowledge of sequential dependencies between melodic events but which, given exposure to a reasonable corpus of music,

would exhibit similar patterns of melodic expectation to those observed in experiments with human participants (see also Bharucha, 1993). This section contains a summary of the computational model that has been implemented to test this prediction with a focus on the principal characteristics of the representation scheme used, how the statistical model acquires knowledge during training, and how this knowledge is applied in generating expectations. Detailed descriptions of the computational methods used by the statistical model may be found elsewhere (Conklin & Witten, 1995; Pearce, Conklin, & Wiggins, 2005; Pearce & Wiggins, 2004).

THE REPRESENTATION SCHEME

The statistical model takes as its *musical surface* (Jackendoff, 1987) sequences of musical events (placed roughly at the note level), representing the instantiation of a finite number of discrete features or attributes which are given descriptive names appearing henceforth in typewriter font to distinguish them as such. Figure 1 presents the first phrase of a chorale melody in standard music notation and in terms of some of the event attributes used in the present research. An event consists of a number of *basic features* representing its onset time (Onset), duration (Duration), and

pitch (Pitch) as shown in the upper panel of Figure 1. In addition, events are associated with basic features representing the current time signature, key signature, mode, and phrase boundaries. These features are derived directly from the score and, in the case of phrase boundaries, the form of the text; it is assumed that these score-based features are representative of perceived features and that the cognitive tasks of melodic segmentation (e.g., Deliège, 1987; Ferrand, Nelson, & Wiggins, 2003), tonality induction (Vos, 2000), and meter induction (e.g., Eck, 2002; Toiviainen & Eerola, 2004) may be addressed independently from the present modeling concerns. Basic features are associated with an alphabet: a finite set of symbols determining the possible instantiations of that feature in a concrete event.

The representation scheme also allows for the derivation of features not present in the basic musical surface but which can be computed from the values of one or more basic features. Examples of such *derived features* include interonset interval (IOI), pitch interval (Interval), contour (Contour), and scale degree (ScaleDegree) as shown in the second panel of Figure 1. In some locations in a melody, a given derived feature may be undefined (denoted by the symbol ⊥) as is the case for Interval for the first event of a melody. In addition, *threaded features* represent a

Onset	0	24	48	72	96	120	144
Duration	24	24	24	24	24	24	48
Pitch	71	71	71	74	72	72	71
IOI	⊥	24	24	24	24	24	24
Interval	⊥	0	0	3	-2	0	-1
Contour	⊥	0	0	1	-1	0	-1
ScaleDegree	4	4	4	7	5	5	4
ThreadBar	⊥	⊥	⊥	⊥	1	⊥	⊥
Interval ⊗ IOI	⊥	(0 24)	(0 24)	(3 24)	(-2 24)	(0 24)	(-1 24)

FIG. 1. The first phrase of the melody from Chorale 151 (Riemenschneider, 1941; see also Figure 5) represented in terms of some of the basic, derived, threaded, and linked features used in the experiments; the symbol ⊥ indicates that a feature is undefined at a given location.

melody in terms of the properties of potentially noncontiguous events. An example is the feature `ThreadBar`, shown in the third panel of Figure 1, which represents the pitch interval between the first events in consecutive bars; at all other metric positions, the feature is undefined. Table 1 summarizes the basic and derived features used in the present research.

Finally, the framework supports the representation of melodies in terms of interactions between primitive features using *linked features*. The linking of *n* features, denoted using the symbol ⊗, simply results in a linked feature whose elements are *n*-tuples composed of the elements of the component features. As an example, the bottom panel of Figure 1 demonstrates the representation of joint melodic and rhythmic structure in a link between pitch interval and interonset interval (`Interval ⊗ IOI`). A linked feature is undefined if any of its component features are undefined at a given location.

Although the present research uses data derived from scores, the representation scheme is rather flexible

and could be extended to represent expressive aspects of music performance (without otherwise changing the nature of the computational model) to the extent that the expressive features of interest can be represented as discrete properties of discrete events or can be derived from representational primitives that can be represented in such a way. We expect this to be the case for the forms of expressive variation in timing and dynamics most commonly studied in the literature on music performance (see C. Palmer, 1997, for a review). The representation scheme is described in full elsewhere (Conklin & Witten, 1995; Pearce et al., 2005) and has been extended to accommodate the representation of homophonic and polyphonic music (Conklin, 2002).

THE MODELING STRATEGY

The computational system itself is based on *n*-gram models commonly used in statistical language modeling (Manning & Schütze, 1999). An *n*-gram is a sequence of *n* symbols and an *n*-gram model is simply a collection of such sequences each of which is associated with a frequency count. During the *training* of the statistical model, these counts are acquired through an analysis of some corpus of sequences (the training set) in the target domain. When the trained model is exposed to a sequence drawn from the target domain, it uses the frequency counts associated with *n*-grams to estimate a probability distribution governing the identity of the next symbol in the sequence given the $n - 1$ preceding symbols. The quantity $n - 1$ is known as the *order* of the model and represents the number of symbols making up the context within which a prediction is made.

The most elementary *n*-gram model of melodic pitch structure (a monogram model where $n = 1$) simply tabulates the frequency of occurrence for each chromatic pitch encountered in a traversal of each melody in the training set. During prediction, the expectations of the model are governed by a zeroth-order pitch distribution derived from the frequency counts and do not depend on the preceding context of the melody. In a digram model (where $n = 2$), however, frequency counts are maintained for sequences of two pitch symbols and predictions are governed by a first-order pitch distribution derived from the frequency counts associated with only those digrams whose initial pitch symbol matches the final pitch symbol in the melodic context.

Fixed-order models such as these suffer from a number of problems. Low-order models (such as the monogram model discussed above) clearly fail to provide an adequate account of the structural influence

TABLE 1. The features used in the present research.

Feature	Description
Pitch	Chromatic pitch
Onset	Onset time
Duration	Duration
IOI	interonset interval
DurRatio	Duration ratio
FirstBar	Whether an event is the first in the current bar
PitchClass	Octave equivalent pitch class or chroma
Interval	Chromatic pitch interval in semitones
IntervalClass	Octave equivalent pitch interval class
Contour	Melodic contour or registral direction
IntFirstPiece	Interval in semitones from the first event in the piece
IntFirstBar	Interval in semitones from the first event in the current bar
IntFirstPhrase	Interval in semitones from the first event in the current phrase
ScaleDegree	Scale degree of chromatic scale constructed on the tonic
InScale	Whether a tone is diatonic in the current key
ThreadTactus	Interval in semitones between events occurring on tactus pulses
ThreadBar	Interval in semitones between the initial events of successive bars
ThreadInitPhr	Interval in semitones between the initial events of successive phrases
ThreadFinalPhr	Interval in semitones between the final events of successive phrases

of the context on expectations. However, increasing the order can prevent the model from capturing much of the statistical regularity present in the training set. An extreme case occurs when the model encounters an *n*-gram that does not appear in the training set in which case it returns an estimated probability of zero. In order to address these problems, the models used in the present research maintain frequency counts during training for *n*-grams of all possible values of *n* in any given context. During prediction, distributions are estimated using a weighted linear combination of all models below a variable order bound, which is determined in each predictive context using simple heuristics designed to minimize model uncertainty. The combination is designed such that higher-order predictions (which are more specific to the context) receive greater weighting than lower-order predictions (which are more general). In a given melodic context, therefore, the predictions of the model may reflect the influence of both the digram model and (to a lesser extent) the monogram model discussed above. Furthermore, in addition to the general, low-order statistical regularities captured by these models, the predictions of the model can also reflect higher-order regularities which are more specific to the current melodic context (to the extent that these exist in the training set). Pearce and Wiggins (2004) give a comprehensive account of the generation of predictions from the trained models, the details of which lie beyond the scope of the present article.

INFERENCE OVER MULTIPLE FEATURES

One final issue to be covered regards the manner in which the statistical model exploits the representation of multiple features of the musical surface described above. The modeling process begins with the selection of a set of features of interest and the training of distinct *n*-gram models for each of these features. For each event in a melody, each feature is predicted using two models: first, the *long-term* model that was trained over the entire training set in the previous step; and second, a *short-term* model that is trained incrementally for each individual melody being predicted.

The task of combining the predictions from all these models is achieved in two stages, both of which use a weighted multiplicative combination scheme in which greater weights are assigned to models whose predictions are associated with lower entropy (or uncertainty) at that point in the melody. In this scheme, a combined distribution is achieved by taking the product of the weighted probability estimates returned by each model for each possible value of the pitch of the next event and then normalizing such that the com-

bined estimates sum to unity over the pitch alphabet. The entropy-based weighting method and the use of a multiplicative as opposed to a linear combination scheme both improve the performance of the model in predicting unseen melodies (Pearce et al., 2005; Pearce & Wiggins, 2004).

In the first stage of model combination, the predictions of models for different features are combined for the long-term and short-term models separately. Distributions from models of derived features are first converted into distributions over the alphabet of the basic feature from which they are derived (e.g., `Pitch`). If a feature is undefined at a given location in a melody, a model of that feature will not contribute to the predictions of the overall system at that location. In the second stage, the two combined distributions (long-term and short-term) resulting from the first step are combined into a single distribution which represents the overall system's final expectations regarding the pitch of the next event in the melody. The use of long- and short-term models is intended to reflect the influences on expectation of both existing extraopus and incrementally increasing intraopus knowledge while the use of multiple features is intended to reflect the influence of regularities in many dimensions of the musical surface. Pearce et al. (2005) give a full technical description of the combination of predictions from models of different melodic features.

Experimental Methodology

The present research has two primary objectives which, in accordance with the level at which the theory is presented (and the manner in which it diverges from the IR theory), are stated at a rather high level of description. The first objective is to test the hypothesis that the statistical model presented above is able to account for the patterns of melodic expectation observed in experiments with human listeners at least as well as quantitative formulations of the IR theory. Since the statistical model acquires its knowledge of sequential melodic structure purely through exposure to music, corroboration of the hypothesis would demonstrate that it is not necessary to posit innate and universal musical rules to account for the observed patterns of melodic expectation; melodic expectancy can be accounted for wholly in terms of statistical induction of both intraopus and extraopus regularities in existing musical corpora.

The methodological approach followed in examining this hypothesis compares the patterns of melodic expectation generated by the computational model to those of human participants observed in previously

reported experiments. Three experiments are presented which elicit expectations in increasingly complex melodic contexts: first, in the context of the single intervals used by Cuddy and Lunny (1995); second, in the context of the excerpts from British folk songs used by Schellenberg (1996) and Krumhansl (1995b); and third, throughout the two chorale melodies used by Manzara, Witten, and James (1992).

In each experiment, the statistical models are compared to the two-factor model of Schellenberg (1997) plus a tonality predictor. Although the two-factor model did not perform as well as that of Krumhansl (1995a) in accounting for the expectations of the listeners in the experiments of Krumhansl et al. (1999, 2000), the converse was true in the experiments of Schellenberg et al. (2002). While the debate surrounding the precise quantitative formulation of the bottom-up system appears likely to continue, this particular IR variant was chosen from those reviewed above because it provides the most parsimonious formulation of the bottom-up principles without loss of predictive power in accounting for the data collected by Cuddy and Lunny (1995) and Schellenberg (1996), which are used in Experiments 1 and 2 respectively. Following common practice, the two-factor model of expectancy is supplemented with a tonality predictor developed in previous research. In the first experiment, the influence of tonality was modeled using the tonal region predictor of Krumhansl (1995a) while the second and third experiments used the Krumhansl and Kessler key profiles for the notated key of the context.

Following Cutting et al. (1992) and Schellenberg et al. (2002), the statistical model and the two-factor model of expectancy are compared on the basis of scope, selectivity, and simplicity. Regarding the scope of the two models, since the individual participant data were not available for any of the experiments and the two models are not nested, Williams' t statistic for comparing dependent correlations (Hittner, May, & Silver, 2003; Steiger, 1980) was used to compare the two models in each experiment. It is expected that the relative performance of the statistical model will increase with longer and more realistic melodic contexts. The selectivity of the models was assessed by using each model to predict random patterns of expectation in the context of the experimental stimuli used in each experiment. Finally, with regard to simplicity, we examine the extent to which the statistical model subsumes the function of the bottom-up components of the two-factor model in accounting for the behavioral data used in each experiment. An alpha level of .05 is used for all statistical tests.

The second objective is to undertake a preliminary examination of which musical features present in (or simply derivable from) the musical surface afford regularities that are capable of supporting the acquisition of the empirically observed patterns of melodic expectation. In each experiment, hypotheses are presented regarding the specific features expected to afford such regularities. The approach taken to testing these hypotheses has been to select sets of features that maximize the fit between experimentally determined human patterns of expectation and those exhibited by the computational model. This was achieved using a forward stepwise selection algorithm (Aha & Bankert, 1996; Blum & Langley, 1997; Kohavi & John, 1996) which, given an empty set of features, considers on each iteration all single feature additions and deletions from the current feature set, selecting the addition or (preferably) deletion that yields the most improvement in the performance metric and terminating when no such addition or deletion yields an improvement.

While this hill-climbing algorithm significantly reduces the size of the effective search space, the solution is not guaranteed to be globally optimal. The use of forward selection and the preference for feature deletions over additions may be justified by the observation that simplicity appears to be a powerful and general organizing principle in perception and cognition (Chater, 1999; Chater & Vitányi, 2003). The performance metrics and feature sets used are described in greater detail separately for each experiment in turn.

All models were trained using a corpus consisting of 152 Canadian folk songs and ballads (Creighton, 1966), 185 of the chorale melodies harmonized by J. S. Bach (Riemenschneider, 1941), and 566 German folk songs (dataset fink) from the Essen Folk Song Collection (Schaffrath, 1992, 1994, 1995). The first of these datasets was obtained from the *Music Cognition Laboratory* at Ohio State University (see http://kern.humdrum.net) while the remaining two were obtained from the *Center for Computer Assisted Research in the Humanities* (CCARH) at Stanford University (see http://www.ccarh.org). Table 2 contains more detailed information about the three datasets, which were selected to represent a range of styles of melodic music within the Western tonal tradition.

In discussing the experimental results, we shall talk about finding support for the influence of a particular feature on melodic expectancy. It should be kept in mind that this shorthand is intended to convey that support has been found for the existence of statistical regularities in a given melodic dimension that increase

TABLE 2. The melodic datasets used for model training.

Description	No. compositions	No. events	Mean events/composition
Canadian folk songs/ballads	152	8,553	56.27
Chorale melodies	185	9,227	49.88
German folk songs	566	33,087	58.46
Total	903	50,867	56.33

TABLE 3. The melodic contexts used in experiment 1 (after Cuddy & Lunny, 1995, Table 2).

Context interval		Second tone	
Interval	Direction	C	F♯
Major second	Ascending	B♭$_3$–C$_4$	E$_4$–F♯$_4$
	Descending	D$_4$–C$_4$	G♯$_4$–F♯$_4$
Minor third	Ascending	A$_3$–C$_4$	D♯$_4$–F♯$_4$
	Descending	E♭$_4$–C$_4$	A$_4$–F♯$_4$
Major sixth	Ascending	E♭$_3$–C$_4$	A$_3$–F♯$_4$
	Descending	A$_4$–C$_4$	D♯$_5$–F♯$_4$
Minor seventh	Ascending	D$_3$–C$_4$	G♯$_3$–F♯$_4$
	Descending	B♭$_4$–C$_4$	E$_5$–F♯$_4$

the fit between the behavior of the statistical model and the observed human behavior.

Experiment 1

Method

The objective in this experiment was to examine how well the statistical model accounts for patterns of expectation following single interval contexts. Cuddy and Lunny (1995) report an experiment in which listeners were asked to rate continuation tones following a two-tone context. The participants were 24 undergraduate students at Queen's University in Canada of whom half were musically trained and half untrained. The stimuli consisted of eight implicative contexts corresponding to ascending and descending intervals of a major second, a minor third, a major sixth, and a minor seventh. All participants heard half of the contexts ending on C$_4$ and half ending on F♯$_4$ (see Table 3) in an attempt to discourage them from developing an overall top-down sense of tonality for the entire experiment. Continuation tones consisted of all 25 chromatic tones from one octave below to one octave above the second tone of the implicative context. The two tones of each context were presented as a dotted half note followed by a quarter note while all continuation tones had a half note duration. These durations were chosen to create a sense of 4/4 meter continuing from the first bar (containing the implicative interval) to the second bar (containing the continuation tone).

The participants were asked to rate how well the continuation tone continued the melody on a scale from 1 (extremely bad continuation) to 7 (extremely good continuation). The experiment yielded 200 continuation tone ratings for each participant. An analysis of variance with the factors music training, context interval, and continuation tone yielded one significant interaction between context interval and continuation tone. Since there was no effect of training and the data exhibited high interparticipant correlation, the ratings for each continuation tone were averaged across participants and training levels. The mean continuation tone ratings for trained and untrained participants are available in Cuddy and Lunny (1995, Appendix).

In the present experiment, the trained model was exposed to each of the eight contexts used by Cuddy and Lunny (1995) for all of which the second tone was F♯$_4$. Due to the short contexts involved, the short-term model was not used in this experiment. In each case, the model returns a single probability distribution (regardless of the number of features it considers) over the set of 25 chromatic pitches ranging from F♯$_3$ to F♯$_5$. Since the distributions returned by the model are

constrained to sum to one and are likely to violate the parametric normality assumption, each of the pitches was assigned a rank according to its estimated probability in inverse order (such that high probability pitches were assigned high ranks). The regression coefficient of the mean ratings obtained by Cuddy and Lunny (1995) regressed on the distribution ranks of the model was used as a performance metric in feature selection. In terms of features used, chromatic pitch (`Pitch`) and pitch class or chroma (`PitchClass`; see Shepard, 1982) were included although they were not expected to exert significant influences on expectancy as a result of the limited context. It was hypothesized that more abstract melodic features such as chromatic pitch interval (`Interval`) and interval class (`IntervalClass`) would be the most important source of regularities underlying melodic expectancy (Dowling & Bartlett, 1981). Pitch contour (`Contour`) was also included to examine the effects of a still more abstract representation of registral direction (Dowling, 1994). It was also hypothesized that the patterns of expectation may reflect a mode of perception in which subsequent tones are appraised in relation to the first tone in the context (`IntFirstPiece`). Given the impoverished context, a sense of tonality may have been inferred based on the first tone of the context as tonic (Cohen, 2000; Cuddy & Lunny, 1995; Longuet-Higgins & Steedman, 1971; Thompson et al., 1997). In spite of the limited context, it

was also hypothesized that pitch may have interacted with rhythmic dimensions of the contexts in the generation of expectations (Jones, 1987; Jones & Boltz, 1989). Consequently, a set of linked features was included in the experiment which represent interactions between three simple pitch-based features (`Pitch`, `Interval`, and `Contour`) and three rhythmic features (`Duration`, `DurRatio`, and `IOI`).

Results

The final set of features selected in this experiment enabled the statistical model to account for approximately 72% of the variance in the mean continuation tone ratings, $R = .85$, $R^2_{adj} = .72$, $F(1,198) = 500.2$, $p < .01$. The relationship between the patterns of expectation exhibited by the model and by the participants in the experiments of Cuddy and Lunny (1995) is plotted with the fitted regression line in Figure 2. The statistical model provided a slightly closer fit to the data than the two-factor model, which accounted for approximately 68% of the variance in the data, $R = .83$, $R^2_{adj} = .68$, $F(3,196) = 141.2$, $p < .01$, although the difference was found not to be significant, $t(197) = 1.1$, $p = .27$.

In order to examine the hypothesis that the statistical model subsumes the function of the bottom-up components of the two-factor model, a more detailed
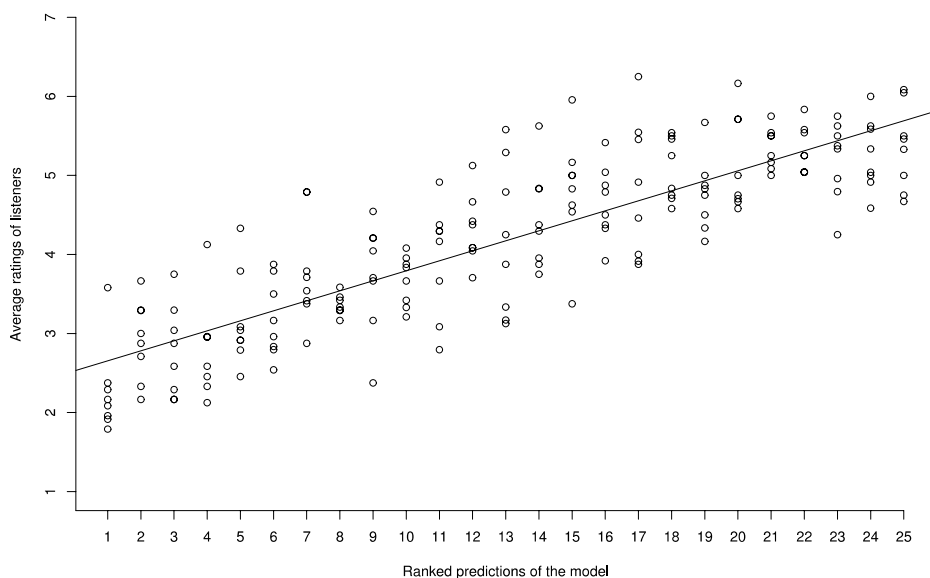

FIG. 2. Correlation between participants' mean goodness-of-fit ratings and the predictions of the statistical model for continuation tones in the experiments of Cuddy and Lunny (1995).

comparison of the two models was conducted. The expectations of the statistical model exhibit significant correlations in the expected directions with both components of the two-factor model: Proximity, $r(198) = -.67$, $p < .01$; and Reversal, $r(198) = .31$, $p < .01$. Furthermore, the fit of the statistical model to the behavioral data was not significantly improved by adding Proximity, $F(1,197) = 1.54$, $p = .22$, Reversal, $F(1,197) < 0.01$, $p = .98$, or both of these factors, $F(2,196) = 0.81$, $p = .45$, to the regression model. This analysis indicates that the statistical model entirely subsumes the function of Proximity and Reversal in accounting for the data collected by Cuddy and Lunny (1995).

Finally, in order to examine the selectivity of the two models, 50 sets of ratings for the stimuli ($N = 200$ in each set) were generated through random sampling from a normal distribution with a mean and standard deviation equivalent to those of the listeners' ratings. With an alpha level of .05, just two of the 50 random vectors were fitted at a statistically significant level by each of the models and there was no significant difference between the fit of the two models for any of the 50 trials. Neither model is broad enough in its scope to successfully account for random data.

The results of feature selection are shown in Table 4. As predicted on the basis of the short contexts, the features selected tended to be based on pitch interval structure. The limited context for the stimulation of expectancy is probably insufficient for the evocation of statistical regularities in chromatic pitch structure. The fact that `Interval ⊗ Duration` was selected over and above its primitive counterpart (`Interval`) suggests that expectations were influenced by the interaction of regularities in pitch interval and duration. It might appear surprising that regularities in rhythmic structure should influence expectations with contexts

so short. Although this may be an artifact, recall that Cuddy and Lunny (1995) carefully designed the rhythmic structure of their stimuli to induce a particular metric interpretation. The issue could be investigated further by systematically varying the rhythmic structure of the stimuli used to obtain goodness-of-fit ratings. Finally, the results reveal a strong influence of `IntFirstPiece` on expectancy which may be partly accounted for by the brevity of the contexts, which do not contain enough information to reliably induce a tonality, combined with the relatively long duration of the first tone. Regularities in the three selected dimensions of existing melodies are such that the statistical model provides an equally close fit to the patterns of expectation observed in the experiment of Cuddy and Lunny (1995) as the two-factor model of Schellenberg (1997).

Experiment 2

Method

The objective of this experiment was to extend the approach of Experiment 1 to patterns of expectation observed after longer melodic contexts drawn from an existing musical repertoire. Schellenberg (1996, Experiment 1) reports an experiment in which listeners were asked to rate continuation tones following eight melodic fragments taken from British folk songs (R. Palmer, 1983; Sharp, 1920). The participants were 20 members of the community of Cornell University in the United States of whom half had limited music training and half had moderate music training. Figure 3 shows the eight melodic contexts of which four are in a minor mode and four in a major mode. They were chosen such that they ended on an implicative interval. Four of the fragments end with one of two small intervals (2 or 3 semitones) in ascending and descending forms while the other four end with one of two large intervals (9 or 10 semitones) in ascending and descending forms. Continuation tones consisted of the 15 diatonic tones in a two-octave range centered on the final tone of the melodic context. The participants were asked to rate how well the continuation tone continued the melody on a scale from 1 (extremely bad continuation) to 7 (extremely good continuation). The experiment yielded 120 continuation tone ratings for each participant. Significant interparticipant correlation for all participants warranted the averaging of the data across participants and training levels. The mean continuation tone ratings are available in Schellenberg (1996, Appendix A).

TABLE 4. The results of feature selection in experiment 1 showing features added to the statistical model and regression coefficients (R) between participants' mean goodness-of-fit ratings and the predictions of the model for continuation tones in the experiments of Cuddy and Lunny (1995); the symbol ⊗ represents a link between two component features.

Stage	Feature added	R
1	Interval ⊗ Duration	.77
2	IntFirstPiece	.84
3	IntervalClass	.85

Fragment 1

Fragment 2

Fragment 3

Fragment 4

Fragment 5

Fragment 6

Fragment 7

Fragment 8

FIG. 3. The melodic contexts used in Experiment 2 (after Schellenberg, 1996, Figure 3).

The procedure used in the present experiment was essentially the same as in Experiment 1 except that the statistical model returned distributions over an alphabet consisting of the diatonic tones an octave above and an octave below the final tone of each melodic fragment. Since the melodic fragments were longer, the short-term model was used in this experiment. Several features, corresponding to hypotheses about the musical regularities underlying the observed patterns of expectation, were added to the set used in Experiment 1. In particular, it was hypothesized that melodic expectations might be influenced by tonality and the interaction of pitch with metric features. It should be emphasized once again that these features were taken from the notated score and may not accurately reflect the perception of tonality and metric accent. In the latter case, however, the stimuli were presented to the participants with a subtle pattern of emphasis in intensity

based on the notated time signature (Schellenberg, 1996) in order to clarify the metrical structure (e.g., in the cases of Fragments 5 and 7 in Figure 3 which might otherwise be more naturally perceived in 2/4 meter).

Regarding metric structure, it was hypothesized that expectations might be influenced by regularities in pitch interval between events occurring on metric pulses (`ThreadTactus`) and the interval of a note from the first note in the bar (`IntFirstBar`) respectively reflecting the influence of tactus and bar level metric salience (Jones, 1987). Regarding the effects of perceived tonality (Balzano, 1982; Krumhansl & Kessler, 1982), it was hypothesized that expectations might be influenced by the representation of a melody in terms of scale degree (`ScaleDegree`). The hypothesis underlying the use of this representational dimension is closely related to an argument made by Krumhansl (1990) that the statistical usage of tones in existing musical traditions is the dominant influence on perceived tonal hierarchies (see also Aarden, 2003). The feature `ScaleDegree` was also linked with `Duration`, `DurRatio`, `IOI`, `Interval`, `IntFirstPiece`, and `FirstBar` to investigate the interactions between perceived tonal structure and these dimensions of melodic, metric, and rhythmic structure (Jones, 1987; Jones & Boltz, 1989).

Results

The final set of features selected in this experiment enabled the statistical model to account for approximately 83% of the variance in the mean continuation tone ratings, $R = .91$, $R^2_{adj} = .83$, $F(1,118) = 571.4$, $p < .01$. The relationship between the patterns of expectation exhibited by the model and by the participants in the experiments of Schellenberg (1996) is plotted with the fitted regression line in Figure 4. The statistical model provided a closer fit to the data than the two-factor model, which accounted for approximately 75% of the variance in the data, $R = .87$, $R^2_{adj} = .75$, $F(3,116) = 121.9$, $p < .01$, and this difference was found to be significant, $t(117) = 2.18$, $p = .03$.

In order to examine the hypothesis that the statistical model subsumes the function of the bottom-up components of the two-factor model, a more detailed comparison of the two models was conducted. The expectations of the statistical model exhibit significant correlations in the expected directions with both components of the two-factor model: Proximity, $r(118) = -.74$, $p < .01$; and Reversal, $r(118) = .49$, $p < .01$. Furthermore, the fit of the statistical model to the behavioral data was not significantly improved by adding Proximity, $F(1,117) = 3.86$, $p = .05$, or Reversal,
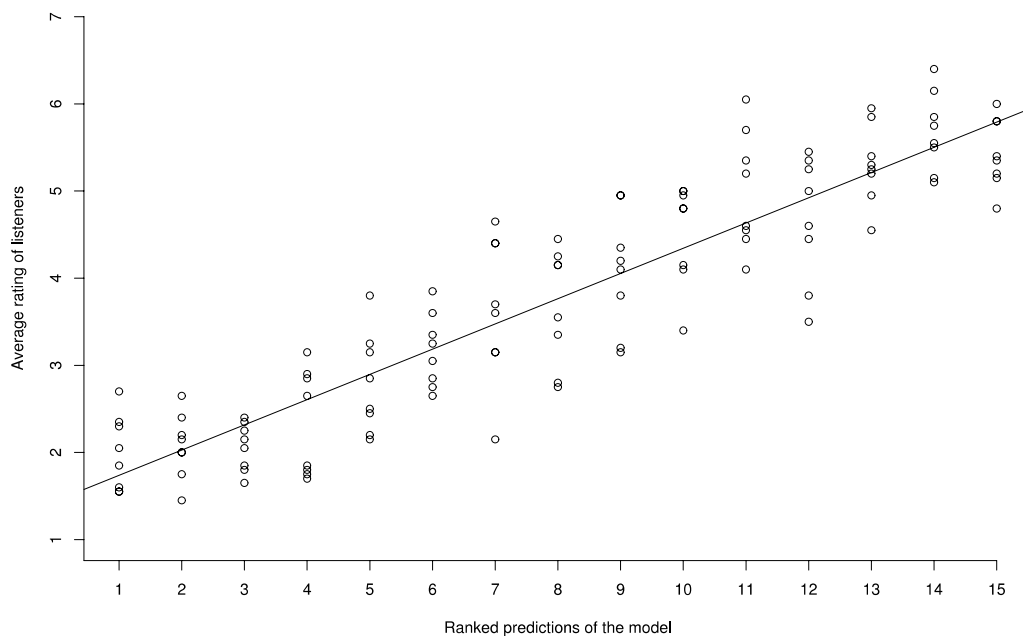

FIG. 4. Correlation between participants' mean goodness-of-fit ratings and the predictions of the statistical model for continuation tones in the experiments of Schellenberg (1996).

$F(1,117) = 1.64$, $p = .2$, to the regression model. However, adding both of these factors did significantly improve the fit of the statistical model to the data, $F(2,116) = 6.03$, $p < .01$. The resulting three-factor regression model accounted for approximately 84% of the variance in the mean continuation tone ratings, $R = .92$, $R^2_{adj} = .84$, $F(3,116) = 210.7$, $p < .01$.

Since the variables of the two-factor model are defined in terms of pitch interval, this departure from the results of Experiment 1 may reflect the relative paucity of features related to pitch interval selected in the present experiment (see Table 5). Since the feature selection algorithm does not cover the space of feature sets exhaustively, it is quite possible that there exist feature sets that include features related to pitch interval, that do not compromise the fit to the data achieved by the present statistical model but for which the addition of the two components of the two-factor model does not yield an improvement. Nonetheless, since the improvement yielded by the addition of the two predictors of the two-factor model was so small (an additional 1% of the variance, given 17% left unaccounted for by the statistical model alone), this analysis indicates that the statistical model *almost* entirely subsumes the function of Proximity and Reversal in accounting for the data collected by Schellenberg (1996).

Finally, in order to examine the selectivity of the two models, 50 sets of ratings for the stimuli ($N = 120$ for each set) were generated through random sampling from a normal distribution with a mean and standard deviation equivalent to those of the listeners' ratings. With an alpha level of .05, just two of the 50 random vectors were fitted at a statistically significant level by each of the models and there was no significant difference between the fit of the

two models for any of the 50 trials. Neither model is broad enough in its scope to successfully account for random data.

The results of feature selection are shown in Table 5. Strong support was found for Pitch especially when linked with IOI, again illustrating the influence of joint regularities in pitch structure and rhythmic structure on expectations. The fact that Pitch was dropped immediately after the addition of Pitch ⊗ IOI suggests not only that the addition of the latter rendered the presence of the former redundant but also that regularities in Pitch, in the absence of rhythmic considerations, provide an inadequate account of the influence of pitch structure on expectations. In contrast to the impoverished contexts used in Experiment 1, the longer contexts used in this experiment are capable of invoking states of expectancy based on regularities in chromatic pitch structure. These regularities are likely to consist primarily of low-order intraopus regularities captured by the short-term model, although potentially higher-order extraopus effects (via the long-term model) may also contribute since two of the training corpora contain Western folk melodies (*cf.* Krumhansl et al., 1999). The features IntFirstBar and IntFirstPiece also contributed to improving the fit of the model to the human data, suggesting that regularities defined in reference to salient events (the first in the piece and the first in the current bar) are capable of exerting strong influences on melodic expectations. Finally, one feature representing a joint influence of regularities in tonal and melodic structure (ScaleDegree ⊗ Interval) was selected. While this feature improved the fit of the statistical model, it is surprising that features modeling tonality were not selected earlier. This may be a result of the fact that British folk melodies are frequently modal (rather than tonal) and the fragments used do not always contain enough information to unambiguously specify the mode (A. Craft, personal communication, September 9, 2003).

Regularities in the four selected dimensions of existing melodies are such that the statistical model provides a closer fit to the patterns of expectation observed in the experiment of Schellenberg (1996) than the two-factor model of Schellenberg (1997).

TABLE 5. The results of feature selection in experiment 2 showing features added to and dropped from the statistical model and regression coefficients (*R*) between participants' mean goodness-of-fit ratings and the predictions of the model for continuation tones in the experiments of Schellenberg (1996); the symbol ⊗ represents a link between two component features.

Stage	Feature added	Feature dropped	R
1	Pitch		.84
2	IntFirstBar		.88
3	IntFirstPiece		.89
4	ScaleDegree ⊗ Interval		.9
5	Pitch ⊗ IOI		.91
6		Pitch	.91

Experiment 3

Method

Most experimental studies of expectancy, including those of Cuddy and Lunny (1995) and Schellenberg (1996), have examined the responses of participants

only at specific points in melodic passages. Results obtained by this method, however, cannot address the question of how expectations change as a melody progresses (Aarden, 2003; Eerola et al., 2002; Schubert, 2001; Toiviainen & Krumhansl, 2003). The purpose of this experiment was to examine the statistical model and the two-factor model (Schellenberg, 1997) in the context of expectations elicited throughout a melodic passage.

Manzara et al. (1992) have used an interesting methodological approach to elicit the expectations of listeners throughout a melody. The goal of their research was to derive an estimate of the entropy of individual pieces within a style according to the predictive models used by human listeners. The experimental stimuli used by Manzara et al. (1992) consisted of the melodies from Chorales 61 and 151 harmonized by J. S. Bach (Riemenschneider, 1941), which are shown in Figure 5. The experimental methodology followed a betting paradigm developed by Cover and King (1978) for estimating the entropy of printed English. Participants interacted with a computer program displaying a score which retained all the information of the original except that the pitch of every note was B_4. Given an initial capital of $S_0 = 1.0$, the participants were asked to move through the score sequentially, selecting the expected pitch of each note and betting a proportion p of their capital repeatedly until the selected pitch was correct, after which they could move to the next note. No time limits were set and the participants could listen to the piece up to and including the current candidate note at any point. At each stage n, the participants' capital was incremented by $20pS_{n-1}$ (there were 20 chromatic pitches to choose from) if the selection was correct and decremented by the proportion bet if it was incorrect. This proportional betting scheme was designed to elicit intuitive probability estimates for the next symbol to be guessed and rewards not only the correct guess but also accurate estimates of the symbol's probability. The entropy of the listener at stage n can be estimated as $\log_2 20 - \log_2 S_n$ where S_n is the capital won by the listener at this stage. Higher entropies indicate greater predictive uncertainty such that the actual pitch of the event is less expected.

Unlike the conventional probe tone method, the betting paradigm allows the collection of responses throughout a melodic passage (but see Toiviainen & Krumhansl, 2003, for a development of the probe tone methodology to allow the collection of real-time continuous responses). In addition, Eerola et al. (2002) report convergent empirical support for the use of entropy as a measure of predictability in melody perception. Furthermore, since it elicits responses prior to revealing the identity of the note and encourages the generation of probability estimates, the betting paradigm offers a more direct measure of expectation than the probe tone method. However, the responses of listeners in the betting paradigm are more likely to reflect the result of conscious reflection than in the probe tone paradigm and may be influenced by a potential learning effect.

The participants in the experiments of Manzara et al. (1992) were grouped into three categories according to formal musical experience: novice, intermediate, and

61: *Jesu Leiden, Pein und Tod* (BWV 159)
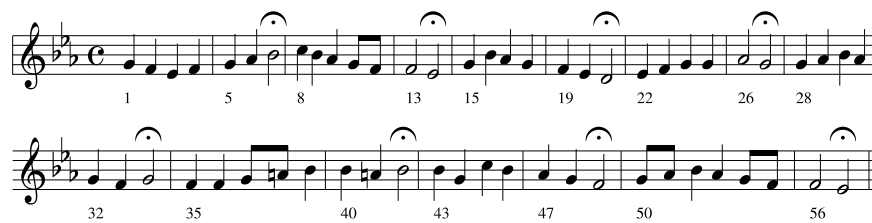
151: *Meinen Jesum laß' ich nicht, Jesus* (BWV 379)

FIG. 5. The two chorale melodies used in Experiment 3 (after Manzara et al., 1992).

expert. The experiment was organized as a competition in two rounds. Five participants in each category took part in the first round with Chorale 151 (see Figure 5), while the two best-performing participants from each category were selected for the second round with Chorale 61 (see Figure 5). As an incentive to perform well, the overall winner in each of the categories won a monetary prize. The capital data for each event were averaged across participants and presented as *entropy profiles* for each chorale melody (see Figures 6 and 7).

Manzara et al. (1992) were able to make some interesting observations about the entropy profiles derived. In particular, it was found that the ultimate tones in phrases tended to be associated with greater predictability than those at the middle and beginning of phrases. High degrees of surprise, on the other hand, were associated with stylistically unusual cadential forms and intervals. The entropy profiles for both pieces also exhibit high uncertainty at the beginning of the piece due to lack of context, followed by increasing predictability as the growing context supported more confident predictions. For both pieces, the results demonstrated a rise in uncertainty near the end of the piece before a steep decline to the final cadence. Witten, Manzara, and Conklin (1994) found a striking similar-

ity between the human entropy profiles and those generated by a statistical model derived from 95 chorale melodies (Conklin & Witten, 1995), suggesting that the relative degrees of expectancy elicited by events throughout the pieces were similar for both the participants and the model.

The experimental procedure used by Manzara et al. (1992) differs from that used by Cuddy and Lunny (1995) and Schellenberg (1996) as does the nature of the data collected. Consequently the methodology followed in this experiment differs slightly from that used in Experiments 1 and 2. The main difference is that the expectations of the statistical model for each note in each melody were represented using entropy (the negative log, base 2, of the estimated probability of the observed pitch). The performance metric was the regression coefficient of the mean entropy estimates for the participants in the experiments of Manzara et al. (1992) regressed on the model entropy. Chorales 61 and 151 were not present in the corpus of chorale melodies used to train the statistical model. Five features were added to the set used in Experiment 2 in order to examine the influence of phrase, metric, and tonal structure on expectations elicited in the longer contexts of the two melodies. Specifically, features were incorpo-
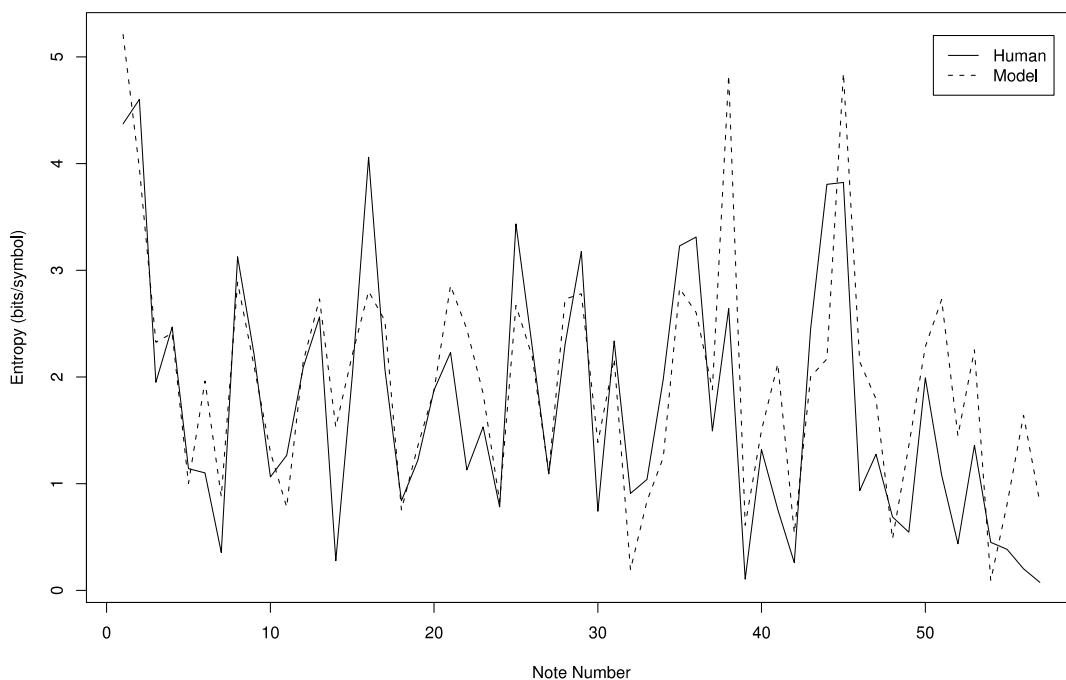

FIG. 6. The entropy profiles for Chorale 61 averaged over participants in the experiment of Manzara et al. (1992) and for the statistical model developed in Experiment 3.
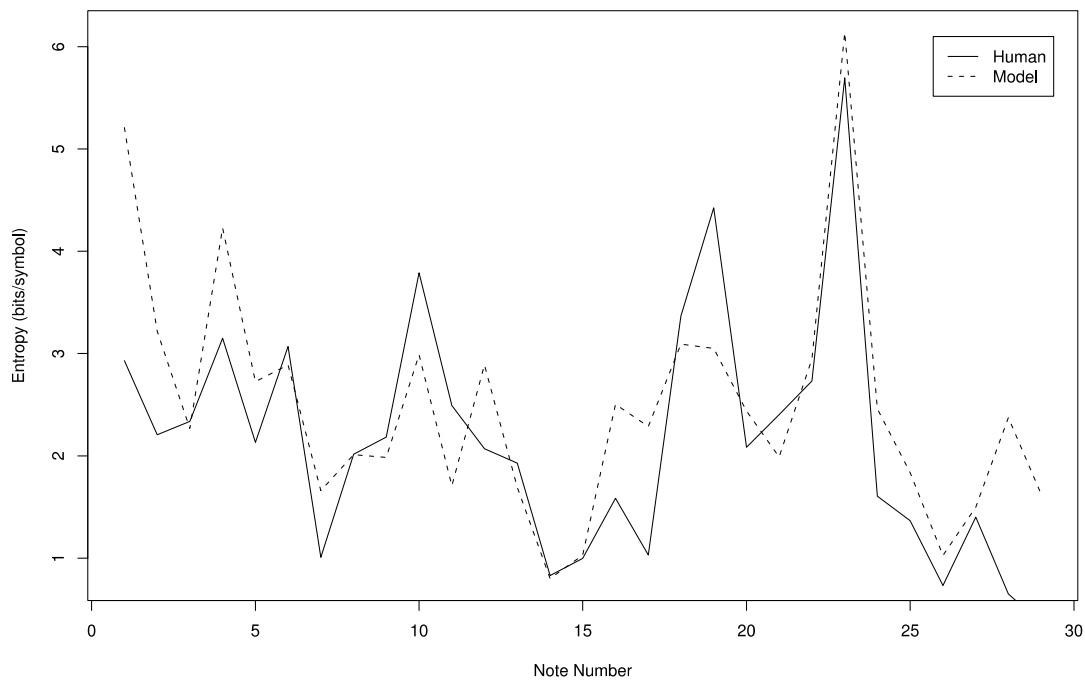
FIG. 7. The entropy profiles for Chorale 151 averaged over participants in the experiment of Manzara et al. (1992) and for the statistical model developed in Experiment 3.

rated to represent pitch interval between the first event in each consecutive bar (`ThreadBar`) and between events beginning or ending consecutive phrases (`ThreadInitPhr` and `ThreadFinalPhr` respectively). A feature representing pitch in relation to the first event in the current phrase (`IntFirstPhrase`) was also added to assess the potential influence of phrase-level salience on expectations. Finally, a feature was added to represent whether or not a tone is diatonic in the notated key of the piece (`InScale`).

Results

The final set of features selected in this experiment enabled the statistical model to account for approximately 63% of the variance in the mean entropy estimates reported by Manzara et al., $R = .8$, $R^2_{adj} = .63$, $F(1,84) = 145$, $p < .01$. Profiles for both model entropy and human entropy are shown in Figures 6 and 7 for Chorales 61 and 151 respectively. The entropy profiles illustrate the close correspondence between model entropy and human entropy throughout each of the chorale melodies (see also Witten et al., 1994). The statistical model provided a closer fit to the data than the two-factor model, which accounted for approximately 13% of the variance in the data, $R = .41$, $R^2_{adj} = .13$,

$F(3,78) = 5.17$, $p < .01$, and this difference was found to be significant, $t(79) = 5.15$, $p < .01$. In the multiple regression analysis of the two-factor model and in comparing it to the statistical model, the data for the first two events of each melody were not used since the two-factor model requires a context of a single interval in order to generate expectations.

In order to examine the hypothesis that the statistical model subsumes the function of the bottom-up components of the two-factor model, a more detailed comparison of the two models was conducted. The expectations of the statistical model exhibit a significant correlation in the expected direction with the Proximity component of the two-factor model, $r(80) = -.41$, $p < .01$, but not with Reversal, $r(80) = .1$, $p = .39$. Furthermore, the fit of the statistical model to the behavioral data was not significantly improved by adding Proximity, $F(1,79) = 0.01$, $p = .91$, Reversal, $F(1,79) = 0.07$, $p = .79$, or both of these factors $F(2,78) = 0.05$, $p = .95$, to the regression model. On this evidence, the statistical model entirely subsumes the function of Proximity and Reversal in accounting for the data collected by Manzara et al. (1992).

Finally, in order to examine the selectivity of the two models, 50 sets of entropy estimates for the two chorales were generated through random sampling

TABLE 6. The results of feature selection in experiment 3 showing features added to the statistical model, the average entropy of the model (*H*) and regression coefficients (*R*) between participants' entropy estimates and those of the model for the two chorale melodies used in the experiments of Manzara et al. (1992); the symbol ⊗ represents a link between two component features.

Stage	Feature added	R	H
1	`IntFirstPiece`	.74	2.29
2	`ScaleDegree ⊗ DurRatio`	.79	2.16
3	`ThreadInitPhr`	.8	2.14

TABLE 7. The results of feature selection for reduced entropy over chorales 61 and 151 in experiment 3 showing features added to the model, the average entropy of the statistical model (*H*) and regression coefficients (*R*) between participants' entropy estimates and those of the model for the two chorale melodies used in the experiments of Manzara et al. (1992); the symbol ⊗ represents a link between two component features.

Stage	Feature added	R	H
1	`ScaleDegree ⊗ Interval`	.66	2.06
2	`Interval ⊗ Duration`	.69	1.97
3	`IntFirstPiece`	.74	1.94
4	`ScaleDegree ⊗ FirstBar`	.75	1.92
5	`ThreadInitPhr`	.76	1.9

from a normal distribution with a mean and standard deviation equivalent to those of the listeners' entropy estimates. With an alpha level of .05, just two of the 50 random vectors were fitted at a statistically significant level by the two-factor model and in only one of these trials was there a significant difference between the fit of the two models. Neither model is broad enough in its scope to successfully account for random data.

The results of feature selection are shown in Table 6. As in Experiments 1 and 2, the feature `IntFirst-Piece` made a strong contribution to the fit of the statistical model. Support was also found for one linked feature representing the influence of tonality (`ScaleDegree ⊗ DurRatio`), and the fact that this feature was selected over its primitive counterpart again provides evidence for the interactive influence of rhythmic and pitch structure on expectancy. Finally, some support was found for an influence of phrase-level regularities on expectancy (`ThreadInitPhr`).

In addition to showing the regression coefficient (*R*), which was used as the evaluation metric in the feature selection experiment, Table 6 also shows the entropy of the statistical model averaged over all events considered during prediction of the two melodies (*H*). The observation that *H* decreases as *R* increases suggests a rational cognitive basis for the selection of melodic features in the generation of expectations: Features may be selected to increase the perceived likelihood (or expectedness) of events and reduce redundancy of encoding (Chater, 1996, 1999). In order to examine this hypothesis, a further selection experiment was run in which features were selected to minimize model uncertainty (as measured by mean per-event entropy) over Chorales 61 and 151. The results of this experiment are shown in Table 7, which shows average model uncertainty (*H*) and the regression coefficient (*R*) of the mean entropy estimates of the participants in the experiments

of Manzara et al. (1992) regressed on the model entropy for each selected system.

Once again, the feature selection results generally exhibit an inverse trend between *R* and *H*. However, while the systems depicted in Tables 6 and 7 show a degree of overlap, Table 7 also reveals that exploiting regularities in certain features (especially those related to melodic interval structure) improves prediction performance but does not yield as close a fit to the behavioral data as the system shown in Table 6. A closer inspection of all 247 systems considered in this experiment revealed a significant negative correlation between *R* and *H* for values of *H* greater than 2.3 bits/symbol $r_s(N = 45) = -.85$, $p < .01$, but not below this point $r_s(N = 202) = -.05$, $p = .46$. If listeners do focus on representations that maximize the perceived likelihood of events, this relationship may be subject to other constraints such as the number and kind of representational dimensions to which they can attend concurrently.

Discussion and Conclusions

The first goal of the present research was to examine whether models of melodic expectancy based on statistical learning are capable of accounting for the patterns of expectation observed in empirical behavioral research. The statistical model and the two-factor model of expectancy (Schellenberg, 1997) were compared on the basis of scope, selectivity, and simplicity (Cutting et al., 1992; Schellenberg et al., 2002). The two models could not be distinguished on the basis of selectivity since neither was found to account for random patterns of expectation in any of the three experiments. Regarding the scope of the two models, the results

demonstrate that the statistical model accounted for the behavioral data as well as, or better than, the two-factor model in all three of the reported experiments. Furthermore, the difference between the two models became increasingly apparent when expectations were elicited in the context of longer and more realistic melodic contexts (see also Eerola et al., 2002). Finally, regarding the simplicity of the two models, the results indicate that the statistical model entirely (or almost entirely in the case of Experiment 2) subsumes the function of the principles of Proximity and Reversal (Schellenberg, 1997) in accounting for the expectations of listeners, rendering the inclusion of these rules in an additional system of innate bottom-up predispositions unnecessary.

Altogether, these experimental results demonstrate that patterns of expectation elicited in a range of melodic contexts can be accounted for in terms of the combined influence of sensitivities to certain dimensions of the musical surface, relatively simple learning mechanisms, and the structure of the musical environment. In contrast to one of the central tenets of the IR theory, universal symbolic rules need not be assumed to account for experimentally observed patterns of melodic expectation. The quantitatively formulated bottom-up and top-down principles of the IR models may be viewed as formalized approximations to behavior that emerges as a result of statistical induction of regularities in the musical environment achieved by a single cognitive system (*cf.* Thompson & Stainton, 1998).

The second goal of the present research was to undertake a preliminary examination of the kinds of melodic feature that afford regularities capable of supporting the acquisition of the observed patterns of expectation. In each experiment, only a small number of features (three in Experiments 1 and 3, and four in Experiment 2) were selected by the forward stepwise selection procedure even though the evaluation functions used did not explicitly penalize the number of features used by the statistical model. In all three experiments, it was found that regularities in pitch structure defined in relation to the first note in a melody are capable of exerting strong influences on expectancy. This influence of primacy on perceived salience suggests that the first note in a melody provides a reference point with which subsequent structures are compared in the generation of expectations (Cohen, 2000; Cuddy & Lunny, 1995; Longuet-Higgins & Steedman, 1971; Thompson et al., 1997). Furthermore, the results of all three experiments provide evidence that expectations are influenced by regularities in the interaction of pitch structure and rhythmic structure (see also Jones, 1987; Jones & Boltz, 1989).

In addition, the experimental results suggest that induced regularities in different melodic features may influence expectancy to varying degrees in different contexts. The short contexts in Experiment 1, for example, tended to generate expectations based on regularities in melodic interval structure rather than chromatic pitch structure. In the second experiment, on the other hand, support was found for the influence of chromatic pitch structure as well as metric structure and tonal regularities. Finally, in Experiment 3, support was found for the influence of tonal structure and phrase-level salience on the generation of expectations. These differences suggest that melodic contexts differ in the extent to which they emphasize different features used in cuing attention to salient events. The results of Experiment 3 also provided some evidence for a relationship, across different feature sets, between the predictive uncertainty of the statistical model and its fit to the behavioral data suggesting that, subject to other constraints, listeners employ representations which increase the perceived likelihood of melodic stimuli (Chater, 1996, 1999). The mechanisms by which attention is drawn to different features in different melodic contexts and how regularities in these dimensions influence expectancy is an important topic for future empirical research. Improved methodologies for eliciting and analyzing continuous responses to music (Aarden, 2003; Eerola et al., 2002; Schubert, 2001; Toiviainen & Krumhansl, 2003) will form an important element in this research.

It is important to note that the concrete implementations of the IR theory discussed herein do not reflect the full complexity of the analytical theory of Narmour (1990, 1992). Further development of these models may lead to improved accounts of behavioral data or the explanation of future empirical observations that cannot be accounted for by the statistical model. In the meantime, however, the experimental results provide strong support for the present theory of expectancy in terms of the influence of melodic context on the invocation of learned regularities. In particular, the results confirm that regularities in existing melodic traditions are sufficient to support the acquisition of observed patterns of expectation. According to the theory, expectations will also be subject to the influence of prior musical experience. The present research used a single corpus of training data, and future research should examine this aspect of the theory in greater depth. It would be predicted, for example, that a model exposed to the music of one culture would predict the expectations of people of that culture better than a model trained on the music of another culture and vice versa (see also Castellano et al., 1984).

The theory also predicts that observed patterns of expectation will become increasingly systematic and complex with increasing age and musical exposure (*cf.* Schellenberg et al., 2002). Future research might examine the developmental profile of expectations exhibited by the statistical model as it learns, yielding testable predictions about developmental trajectories in the acquisition of melodic expectations exhibited by infants (see also Plunkett & Marchman, 1996).

Another fruitful avenue for future research involves a more detailed examination of the untested assumptions of the statistical model, the elaboration of the theory, and the proposition of hypotheses at finer levels of detail (Desain, Honing, Thienen, & Windsor, 1998). Such hypotheses might concern, for example, the developmental status of the features assumed to be present in the musical surface and the derivation of other features from this surface as well as how the interaction between the long- and short-term models is related to the effects of intraopus and extraopus experience. The examination of expectations for more complex musical structures embedded in polyphonic contexts may reveal inadequacies of the model. For example, its reliance on local context in generating predictions may prove insufficient to account for the perception of nonlocal dependencies and recursively embedded structure (Lerdahl & Jackendoff, 1983). Conversely, the computational model may be overspecified in some regards as a model of human cognition. For example, schematic influences on

expectancy are likely to be subject to the effects of limitations on working memory although the model is not explicitly constrained in this regard (Reis, 1999).

To conclude, not only does the theory put forward in this article provide a compelling account of existing data on melodic expectancy, but it also makes a number of predictions for future research. In this regard, the modeling strategy followed in the present research constitutes a rich source of new hypotheses regarding the influence of musical context and experience on expectations and provides a useful framework for the empirical examination of these hypotheses.

Author Note

Address correspondence to: Marcus T. Pearce, Department of Computing, Goldsmiths College, University of London, New Cross, London SE14 6NW, UK. E-MAIL *m.pearce@gold.ac.uk*

References

AARDEN, B. (2003). *Dynamic melodic expectancy*. Unpublished doctoral dissertation, Ohio State University, Columbus.

AHA, D. W., & BANKERT, R. L. (1996). A comparative evaluation of sequential feature selection algorithms. In D. Fisher & H. J. Lenz (Eds.), *Learning from data: AI and statistics V* (pp. 199–206). New York: Springer.

BALZANO, G. J. (1982). The pitch set as a level of description for studying musical pitch perception. In M. Clynes (Ed.), *Music, mind and brain* (pp. 321–351). New York: Plenum.

BERGESON, T. R. (1999). Melodic expectancy in infancy. *Journal of the Acoustical Society of America*, *106*, 2285.

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

BHARUCHA, J. J. (1987). Music cognition and perceptual facilitation: A connectionist framework. *Music Perception*, *5*, 1–30.

BHARUCHA, J. J. (1993). Tonality and expectation. In R. Aiello (Ed.), *Musical perceptions* (pp. 213–239). Oxford: Oxford University Press.

BHARUCHA, J. J., & STOECKIG, K. (1986). Reaction time and musical expectancy: Priming of chords. *Journal of Experimental Psychology: Human Perception and Performance*, *12*, 403–410.

BLUM, A., & LANGLEY, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, *97*, 245–271.

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

CASTELLANO, M. A., BHARUCHA, J. J., & KRUMHANSL, C. L. (1984). Tonal hierarchies in the music of North India. *Journal of Experimental Psychology: General*, *113*, 394–412.

CHATER, N. (1996). Reconciling simplicity and likelihood principles in perceptual organisation. *Psychological Review*, *103*, 566–581.

CHATER, N. (1999). The search for simplicity: A fundamental cognitive principle? *The Quarterly Journal of Experimental Psychology*, *52A*, 273–302.

CHATER, N., & VITÁNYI, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*, 19–22.

COHEN, A. J. (2000). Development of tonality induction: Plasticity, exposure and training. *Music Perception*, *17*, 437–459.

CONKLIN, D. (2002). Representation and discovery of vertical patterns in music. In C. Anagnostopoulou, M. Ferrand, & A. Smaill (Eds.), *Proceedings of the second international conference of music and artificial intelligence*: *Vol. 2445* (pp. 32–42). Berlin: Springer.

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

COVER, T. M., & KING, R. C. (1978). A convergent gambling estimate of the entropy of English. *IEEE Transactions on Information Theory*, *24*, 413–421.

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

CROSS, I. (1995). Review of *The analysis and cognition of melodic complexity: The implication-realization model*, Narmour (1992). *Music Perception*, *12*, 486–509.

CUDDY, L. L., & LUNNY, C. A. (1995). Expectancies generated by melodic intervals: Perceptual judgements of continuity. *Perception and Psychophysics*, *57*, 451–462.

CUTTING, J. E., BRUNO, N., BRADY, N. P., & MOORE, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgements of perceived depth. *Journal of Experimental Psychology: General*, *121*, 364–381.

DELIÈGE, I. (1987). Grouping conditions in listening to music: An approach to Lerdahl and Jackendoff's grouping preference rules. *Music Perception*, *4*, 325–360.

DESAIN, P., HONING, H., THIENEN, H. VAN, & WINDSOR, L. (1998). Computational modelling of music cognition: Problem or solution. *Music Perception*, *16*, 151–166.

DOWLING, W. J. (1994). Melodic contour in hearing and remembering melodies. In R. Aiello & J. Sloboda (Eds.), *Musical perceptions* (pp. 173–190). Oxford: Oxford University Press.

DOWLING, W. J., & BARTLETT, J. C. (1981). The importance of interval information in long-term memory for melodies. *Psychomusicology*, *1*, 30–49.

ECK, D. (2002). Finding downbeats with a relaxation oscillator. *Psychological Research*, *66*, 18–25.

EEROLA, T. (2004a). Data-driven influences on melodic expectancy: Continuations in North Sami Yoiks rated by South African traditional healers. In S. D. Lipscomb, R. Ashley, R. O. Gjerdingen, & P. Webster (Eds.), *Proceedings of the Eighth International Conference of Music Perception and*

Cognition (pp. 83–87). Adelaide, Australia: Causal Productions.

EEROLA, T. (2004b). *The dynamics of musical expectancy: Cross-cultural and statistical approaches to melodic expectations*. Doctoral dissertation, Faculty of Humanities, University of Jyväskylä, Finland. (Jyväskylä Studies in Humanities, 9)

EEROLA, T., TOIVIAINEN, P., & KRUMHANSL, C. L. (2002). Real-time prediction of melodies: Continuous predictability judgements and dynamic models. In C. Stevens, D. Burnham, E. Schubert, & J. Renwick (Eds.), *Proceedings of the Seventh International Conference on Music Perception and Cognition* (pp. 473–476). Adelaide, Australia: Causal Productions.

ELMAN, J. L., BATES, E. A., JOHNSON, M. H., KARMILOFF-SMITH, A., PARISI, D., & PLUNKETT, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

FERRAND, M., NELSON, P., & WIGGINS, G. (2003). Unsupervised learning of melodic segmentation: A memory-based approach. In R. Kopiez, A. C. Lehmann, & C. Wolf (Eds.), *Proceedings of the 5th Triennial ESCOM Conference* (pp. 141–144). Hanover, Germany: Hanover University of Music and Drama.

GJERDINGEN, R. O. (1999). Apparent motion in music? In N. Griffith & P. M. Todd (Eds.), *Musical networks: Parallel distributed perception and performance* (pp. 141–173). Cambridge, MA: MIT Press/Bradford Books.

HITTNER, J. B., MAY, K., & SILVER, N. C. (2003). A Monte Carlo evaluation of tests for comparing dependent correlations. *The Journal of General Psychology*, *130*, 149–168.

JACKENDOFF, R. (1987). *Consciousness and the computational mind*. Cambridge, MA: MIT Press.

JAYNES, E. T. (2003). *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.

JONES, M. R. (1987). Dynamic pattern structure in music: Recent theory and research. *Perception and Psychophysics*, *41*, 621–634.

JONES, M. R., & BOLTZ, M. G. (1989). Dynamic attending and responses to time. *Psychological Review*, *96*, 459–491.

KESSLER, E. J., HANSEN, C., & SHEPARD, R. N. (1984). Tonal schemata in the perception of music in Bali and the West. *Music Perception*, *2*, 131–165.

KOHAVI, R., & JOHN, G. H. (1996). Wrappers for feature subset selection. *Artificial Intelligence*, *97*, 273–324.

KRUMHANSL, C. L. (1990). *Cognitive foundations of musical pitch*. Oxford: Oxford University Press.

KRUMHANSL, C. L. (1995a). Effects of musical context on similarity and expectancy. *Systematische Musikwissenschaft*, *3*, 211–250.

KRUMHANSL, C. L. (1995b). Music psychology and music theory: Problems and prospects. *Music Theory Spectrum*, *17*, 53–90.

KRUMHANSL, C. L. (1997). Effects of perceptual organisation and musical form on melodic expectancies. In M. Leman (Ed.),

Music, Gestalt and computing: Studies in cognitive systematic musicology (pp. 294–319). Berlin: Springer.

KRUMHANSL, C. L., & KESSLER, E. J. (1982). Tracing the dynamic changes in perceived tonal organisation in a spatial representation of musical keys. *Psychological Review*, *89*, 334–368.

KRUMHANSL, C. L., LOUHIVUORI, J., TOIVIAINEN, P., JÄRVINEN, T., & EEROLA, T. (1999). Melodic expectation in Finnish spiritual hymns: Convergence of statistical, behavioural and computational approaches. *Music Perception*, *17*, 151–195.

KRUMHANSL, C. L., TOIVANEN, P., EEROLA, T., TOIVIAINEN, P., JÄRVINEN, T., & LOUHIVUORI, J. (2000). Cross-cultural music cognition: Cognitive methodology applied to North Sami yoiks. *Cognition*, *76*, 13–58.

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

LONGUET-HIGGINS, H. C., & STEEDMAN, M. J. (1971). On interpreting Bach. In B. Meltzer & D. Michie (Eds.), *Machine intelligence 6* (pp. 221–241). Edinburgh, UK: Edinburgh University Press.

MACKAY, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press.

MANNING, C. D., & SCHÜTZE, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

MANZARA, L. C., WITTEN, I. H., & JAMES, M. (1992). On the entropy of music: An experiment with Bach chorale melodies. *Leonardo*, *2*, 81–88.

MARR, D. (1982). *Vision*. San Francisco: W. H. Freeman.

MCCLAMROCK, R. (1991). Marr's three levels: A re-evaluation. *Minds and Machines*, *1*, 185–196.

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

MEYER, L. B. (1973). *Explaining music: Essays and explorations*. Chicago: University of Chicago Press.

MITCHELL, T. M. (1997). *Machine learning*. New York: McGraw Hill.

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

NARMOUR, E. (1991). The top-down and bottom-up systems of musical implication: Building on Meyer's theory of emotional syntax. *Music Perception*, *9*, 1–26.

NARMOUR, E. (1992). *The analysis and cognition of melodic complexity: The implication-realisation model*. Chicago: University of Chicago Press.

NARMOUR, E. (1999). Hierarchical expectation and musical style. In D. Deutsch (Ed.), *The psychology of music* (2nd ed., pp. 441–472). New York: Academic Press.

NOLAN, D. (1997). Quantitative parsimony. *British Journal for the Philosophy of Science*, *48*, 329–343.

ORAM, N., & CUDDY, L. L. (1995). Responsiveness of Western adults to pitch-distributional information in melodic sequences. *Psychological Research*, *57*, 103–118.

PALMER, C. (1997). Music performance. *Annual Review of Psychology*, *48*, 115–138.

PALMER, R. (Ed.). (1983). *Folk songs collected by Ralph Vaughan Williams*. London: Dent.

PAUL, G. (1993). Approaches to abductive reasoning: An overview. *Artificial Intelligence Review*, *7*, 109–152.

PEARCE, M. T., CONKLIN, D., & WIGGINS, G. A. (2005). Methods for combining statistical models of music. In U. K. Wiil (Ed.), *Computer music modelling and retrieval* (pp. 295–312). Heidelberg, Germany: Springer Verlag.

PEARCE, M. T., & WIGGINS, G. A. (2004). Improved methods for statistical modelling of monophonic music. *Journal of New Music Research*, *33*, 367–385.

PLUNKETT, K., & MARCHMAN, V. (1996). Learning from a connectionist model of the acquisition of the past tense. *Cognition*, *61*, 299–308.

POPPER, K. (1959). *The logic of scientific discovery*. London: Hutchinson and Co.

POVEL, D. J., & JANSEN, E. (2002). Harmonic factors in the perception of tonal melodies. *Music Perception*, *20*, 51–85.

REIS, B. Y. (1999). *Simulating music learning with autonomous listening agents: Entropy, ambiguity and context*. Unpublished doctoral dissertation, Computer Laboratory, University of Cambridge, UK.

RIEMENSCHNEIDER, A. (1941). *371 harmonised chorales and 69 chorale melodies with figured bass*. New York: G. Schirmer.

RUSSO, F. A., & CUDDY, L. L. (1999, March). *A common origin for vocal accuracy and melodic expectancy: Vocal constraints*. Paper presented at the Joint Meeting of the Acoustical Society of America and the European Acoustics Association, Berlin, Germany. (Published in *Journal of the Acoustical Society of America, 105*, 1217)

SAFFRAN, J. R., JOHNSON, E. K., ASLIN, R. N., & NEWPORT, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*, 27–52.

SCHAFFRATH, H. (1992). The ESAC databases and MAPPET software. *Computing in Musicology*, *8*, 66.

SCHAFFRATH, H. (1994). The ESAC electronic songbooks. *Computing in Musicology*, *9*, 78.

SCHAFFRATH, H. (1995). The Essen folksong collection. In D. Huron (Ed.), *Database containing 6,255 folksong transcriptions in the Kern format and a 34-page research guide* [computer database]. Menlo Park, CA: CCARH.

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

SCHELLENBERG, E. G. (1997). Simplifying the implication-realisation model of melodic expectancy. *Music Perception*, *14*, 295–318.

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

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

Schmuckler, M. A. (1990). The performance of global expectations. *Psychomusicology*, *9*, 122–147.

Schmuckler, M. A. (1997). Expectancy effects in memory for melodies. *Canadian Journal of Experimental Psychology*, *51*, 292–305.

Schubert, E. (2001). Continuous measurement of self-report emotional responses to music. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and emotion* (pp. 393–414). Oxford: Oxford University Press.

Sharp, C. J. (Ed.). (1920). *English folk songs* (Vols. 1–2, selected edition). London: Novello.

Shepard, R. N. (1982). Structural representations of musical pitch. In D. Deutsch (Ed.), *Psychology of music* (pp. 343–390). New York: Academic Press.

Sober, E. (1981). The principle of parsimony. *British Journal for the Philosophy of Science*, *32*, 145–156.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251.

Thompson, W. F. (1996). Eugene Narmour: *The analysis and cognition of basic musical structures* (1990) and *The analysis and cognition of melodic complexity* (1992): A review and empirical assessment. *Journal of the American Musicological Society*, *49*, 127–145.

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

Thompson, W. F., & Stainton, M. (1996). Using *Humdrum* to analyse melodic structure: An assessment of Narmour's implication-realisation model. *Computing in Musicology*, *12*, 24–33.

Thompson, W. F., & Stainton, M. (1998). Expectancy in Bohemian folk song melodies: Evaluation of implicative principles for implicative and closural intervals. *Music Perception*, *15*, 231–252.

Toiviainen, P., & Eerola, T. (2004). The role of accent periodicities in metre induction: A classification study. In S. D. Lipscomb, R. Ashley, R. O. Gjerdingen, & P. Webster (Eds.), *Proceedings of the Eighth International Conference of Music Perception and Cognition* (pp. 422–425). Adelaide, Australia: Causal Productions.

Toiviainen, P., & Krumhansl, C. L. (2003). Measuring and modelling real-time responses to music: The dynamics of tonality induction. *Perception*, *32*, 741–766.

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

von Hippel, P. T. (2002). Melodic-expectation rules as learned heuristics. In C. Stevens, D. Burnham, E. Schubert, & J. Renwick (Eds.), *Proceedings of the Seventh International Conference on Music Perception and Cognition* (pp. 315–317). Adelaide, Australia: Causal Productions.

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

Vos, P. G. (2000). Tonality induction: Theoretical problems and dilemmas. *Music Perception*, *17*, 403–416.

Vos, P. G., & Pasveer, D. (2002). Goodness ratings of melodic openings and closures. *Perception and Psychophysics*, *64*, 631–639.

Vos, P. G., & Troost, J. M. (1989). Ascending and descending melodic intervals: Statistical findings and their perceptual relevance. *Music Perception*, *6*, 383–396.

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