Learning Melodic Expectancy: Musical Predictability in an SRN

Brandon Abbs (brandon-abbs@uiowa.edu)
Department of Psychology, E11 Seashore Hall
Iowa City, IA 52242 USA

Prahlad Gupta (prahlad-gupta@uiowa.edu)
Department of Psychology, E11 Seashore Hall
Iowa City, IA 52242 USA

Abstract

Descriptive models of music cognition propose various principles as underlying melodic expectancy, however there is very little discussion regarding the processes involved in the acquisition of melodic expectation. To explore the potential role of learning processes, a simple recurrent network (SRN) was trained on a set of musical sequences to examine the degree to which the principles described by Schellenberg's (1997) two-factor model might be learned through musical exposure. The principle of pitch proximity, but not pitch reversal constrained the model's expectations of tones following melodic fragments. Implications for this model in the area of music perception and sequential modeling are discussed, as are potential extensions of this simple system.

Keywords: melodic expectancy; simple recurrent network; learning

Expectancy in Music Cognition

When hearing music, much of the listener's experience depends upon how much a given note is expected given the current musical context. This expectancy may be thought of as a dynamic probability distribution that changes depending on global context variables, such as the style, mode, or key of a musical piece, as well as local context variables, such as the specific notes that have recently been heard (Carlsen, 1981; Krumhansl et al., 1999).

Many investigations in music cognition have determined what musical variables lead to what expectancies in order to understand both the psychology and art of music (see Krumhansl, 1995; 2000 for a review). The focus of the present investigation will be the local context with the goal of investigating how the underlying musical knowledge being probed in these types of investigations may be learned through musical exposure. This goal will be pursued by examining how a connectionist network designed for sequential processing, Elman's (1988) Simple Recurrent Network (SRN), develops melodic expectancy from a small corpus of simple melodies. The novelty of this work in relation to previous models (e.g., Bharucha & Todd, 1989; Krumhansl et al., 1999; Mozer, 1991; Page, 1999) is the model's simplicity and the analysis of the model's performance.

The Implication-Realization (I-R) Model

One model that has gone to great lengths in describing the role of local context in melodic expectancy is the Implication-Realization (I-R) Model (Narmour, 1990). This model de-emphasizes a global, or stylistic, analysis of music and instead focuses on note-to-note relationships to examine how musical implications and realizations are perceived (Narmour, p. ix). According to this model, and other theoretical descriptions of music cognition upon which it was built, an implication is created whenever two notes (an interval) are perceived as incomplete, or open. This implicative interval creates an expectation for the next note of the melody, which is the realization of this interval. The interval between the realized note and the second tone of the implicative interval is the realized interval.

For any given implicative interval, there will be a set of tones that are implied by the interval and the strength of their implication will be graded. That is, certain tones will be strongly implied, while others will be weakly implied, and still others will fall in between these extremes. The goal of theoretical models of music cognition like the I-R model is to determine where individual notes fall along this continuum for different types of implicative intervals and what general principles may be used to describe similar types of expectancies. While an entire volume (Narmour, 1990) has been dedicated to the details of the I-R model, the present discussion will focus on Schellenberg's (1997) simplification of the model into two orthogonal principles.

The Two-Factor Model

One criticism of the full I-R model is that its principles are over specified and overlap with one another (Schellenberg, 1997). In recognition of this fact, and in order to uncover the fundamental components of melodic expectation, Schellenberg advocated a reduction of the I-R model down to two orthogonal principles, pitch proximity and pitch reversal. Apart from being orthogonal, these principles are also dependent on different contextual windows: Pitch proximity considers only the last note of a sequence, but pitch reversal requires the consideration of the last two notes (i.e., the implicative interval) of a sequence.

Pitch Reversal

Pitch reversal collapses the registral return and registral direction principles in the I-R model into one principle. Following a large implicative interval ($\geq$ 7 semitones), the realized interval should either be lateral or in the reverse direction, thus changing the melodic contour established by the interval. Further, with a change in direction, the note should be either a complete return, or a near return. With a small interval, the only contributor to expectation is the return characteristic.

Pitch Proximity

Pitch proximity is essentially identical to the proximity principle of the I-R model in that, the closer a tone is to the previous pitch, the greater the expectation for that tone. The unison tone (i.e., the same tone) is the most expected tone and expectations fall off linearly and absolutely from this tone.

Schellenberg (1997) evaluated the predictive power of the I-R and Two-Factor models by applying them to a set of expectancy ratings reported by Cuddy and Lunney (1995). Expectancy ratings are collected through a melodic expectation task wherein participants are played a melodic fragment (i.e., a melody ending in an implicative interval) followed by a probe tone. The participants are asked to rate the tone according to how well it completes the melodic fragment following the logic that a highly expected tone will be given a high rating. Typically, multiple fragments, each with a unique implicative interval, are used and over the course of the entire experiment these fragments are heard in conjunction with a wide range of probe tones. This procedure results in an expectancy profile for each melody for each subject. A multiple regression is then performed on these ratings using the principles of the models as predictors, with a model's success being determined by its predictive power.

Schellenberg (1997) found that the Two-Factor model explained 72.5% of the variance in the expectancy ratings reported by Cuddy and Lunney (1995), compared to the I-R model's 64.0%. In a prospective study of similar design, Schellenberg et al. (2002) again found the two-factor model to be as successful or more successful than the I-R model in accounting for the variance of the listener's ratings (but see Krumhansl, 1995; Krumhansl et al., 1999; Thompson, Cuddy, & Plaus, 1997). These results raise the question of whether pitch proximity and pitch reversal are the fundamental principles of melodic expectancy. Further, the fact that these principles require different degrees of context, as well other evidence that these principles appear to be sensitive to one's musical culture (Carlsen, 1981; Thompson et al., 1997) raises the question of whether these principles might be learned through musical exposure. The goal here is to test the feasibility of learning through mere exposure and to propose a candidate learning process.

Developing Melodic Expectation

The proximity principle of the two-factor model is predictive of the expectancy ratings of children as young as seven years old as well as those of adults and children of intermediate ages (Schellenberg et al., 2002). Further, this principle explains a majority of the variance as compared to the principle of pitch reversal in all groups. Pitch reversal, on the other hand, only emerges as a significant predictor in adults, or children with an elevated musical ability. For this latter group, pitch reversal emerges in children as young as five years old.

The evidence for the emergence of pitch reversal only after significant musical experience has thus been presented, but the emergence of pitch proximity (as opposed to an innate predisposition) is still in question. Its presence in children as young as seven and five years old may provide evidence against a learning-based account, however this depends largely on how easily such a principle might be learned in a given stimulus environment. It is the goal of the present examination to determine just how easily such a principle might be learned by an SRN exposed to sequences of simplistic melodies, such as those found in nursery rhymes and folk songs. The choice of an SRN for the present project was driven by the fact that it is a computational system with very little structure, yet the ability to process sequences of information and the fundamental characteristic of expectancy in its operation. Further, the SRN's demonstrated success with a wide variety of procedural learning tasks (e.g., syntax, artificial grammar, digit entering, word form, and action learning) links the computational model to a potential learning process that further empirical investigations can aim to verify.

The Computational Model

Architecture

In order to maintain the focus of the present investigation on the note-to-note processing of melodies with notes being described as abstract primitives, and thus maintain the spirit of Narmour's (1990) original analysis, the representation of notes in the network were localist. Specifically, twelve nodes in the input layer and twelve nodes in the output layer represented each of the tones on a chromatic scale (see Figure 1). Taking advantage of the principle of enharmonic equivalence (whereby sharpening a note is perceptually equivalent to flattening a note one whole tone above the sharpened note), one node was used to represent equivalent sharps and flats (e.g., A-sharp/B-flat).

In order to easily expand the representational capability of the network and represent similarity between the same notes in different octaves, octave information was also represented by a single localist node. Four nodes in the input layer and the output layer each represented a different octave according to the number given to them in musical notation (e.g., the octave below middle C=3, the octave containing middle C=4, and so on). Together, these sixteen input and output nodes can represent any note from $C_3 - B_6$.

Inherent within the SRN is the abstract representation of local musical context through recurrence between the hidden layer and a separate context layer, both of

which lie between the input and output layers. The hidden unit sends activation through one-to-one, non-modifiable, connections to the context layer, which in turn sends activation through modifiable connections back to the hidden layer. Upon presentation of a note, the resultant activation at the hidden layer is copied onto the context layer. This context is then presented alongside the next input event and the context's influence is determined by the strength of the connection weights between the context and the hidden layer. Thus, information about previous notes and the current note is available to the model. The hidden layer and context layer were each made up of four units.

Training Corpus

The training set for the model was created from 100 melodic sequences. These sequences were actual songs chosen from three books containing traditional, Western, children's and folk songs (Mitchell, 1968, Raph, 1964; Seeger, 1948). Transcriptions of these songs in the ABC ASCII notation were downloaded from an on-line depository (Chambers, n.d.). 84 of these sequences were completely unique and 16 of these sequences were repeated titles transposed into different keys. The downloaded text files were then processed so as to extract the sequence of tones that make up the melody of the song. In all, 9,175 stimulus events were represented.

Training Procedure

During training, songs from the training corpus were selected randomly without replacement. For each song, notes were presented one at a time to the SRN's input layer, resulting in activation at the output layer, which will be interpreted as the network's prediction, or expectation, of what the next note will be. In order to train the network, the actual next note (or in the case of the last note of a song, a blank vector) was used as a target and the back-propagation learning algorithm was used to modify the connection weights of the system after each stimulus event based on the error between the model's prediction and the actual next note. The learning parameters of this algorithm were set as follows: learning rate = .01, momentum = .9, hysteresis, or μ, = .3. One training epoch constituted one pass through the training set, or one presentation of all one hundred songs.

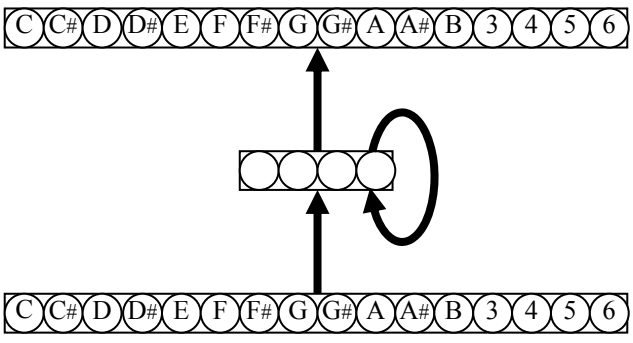

Figure 1: Representational scheme used in the present SRN. The first 12 input and output nodes are localist units representing each note on the chromatic scale. The final 4 nodes are localist units representing each octave region the network might encounter.

Testing Procedure

The testing materials were notations of the melodic fragments presented to participants by Schellenberg et al. (2002, Figure 3). Each fragment was 14 - 16 notes in length and selected from an Acadian (French-speaking Canadians from the Maritime Provinces) folk song. Fragments were selected from songs in the Acadian culture to eliminate the possibility that participants would have any a priori familiarity with the fragments. Further, fragments were selected from these songs such that they ended in an upward implicative interval, which is considered to be more implicative than a downward interval. Lastly, two of the implicative intervals were large (9 & 10 semitones), providing an opportunity for the pitch reversal principle to emerge, while the other two intervals were small (2 & 3 semitones).

The model was tested on every epoch. At test, learning was turned off and each of the melodic fragments was presented to the model. Following the final stimulus event of a fragment (the second tone of the implicative interval), the activation levels of the output nodes were recorded. The model's prediction of a particular note in a specific octave (e.g., C_4) was computed as the dot product of this final output vector and the target vector for the note in question. This dot product was then treated as the expectancy rating of the model and scaled so as to match the likert scale ratings reported by Schellenberg et al. (2002, Figure 4). This measure was used in order to functionally bind note and octave information, which must be done in order to translate the activation of the output vector into a metric akin to an expectancy rating. After this measure was taken for each test sequence, learning was turned back on and the model was trained for another epoch before being tested again.

We did not expect that this simple model, learning from a simple training set, would be able to capture all the nuances in expectancy ratings that human data show. Rather, we wished to examine whether the principles developed as descriptions of these expectancy ratings might emerge as a result of learning through exposure to nursery rhymes. Given the pervasiveness across age groups and explanatory power of pitch proximity in the ratings reported by Schellenberg et al. (2002), this principle seemed the one most likely to be evident in the computational model. Pitch reversal, on the other hand, would only be evident to the extent that a.) the model was able to take larger units of context into account (as proposed by Schellenberg et al. for humans) and b.) large implicative intervals were prevalent in the simple melodies of the training set.

Simulation Results & Analysis

Table 1 shows correlations between the model's expectancy ratings and the behavioral data, estimated from

Figure 4 of Schellenberg et al. (2002). Overall, the model correlates the best with expectancy ratings generated for smaller implicative intervals than for larger implicative intervals and the ratings of older children better than adults and younger children (see Table 1). Furthermore, these correlations are evident after just one epoch of training. For melody 1, the model shows a significant correlation ($p < .05$ for all correlations) with ratings from all age groups [rs (13) = .829, .843, .761 for adults, older children, and younger children respectively]. For melody 2, the model correlates significantly with only the adults ($r = .699$) and the older children ($r = .866$). For melody 3, which ends in a large implicative interval, the model again correlates significantly with adults ($r = .515$) and older children ($r = .602$), but not younger children. Lastly, for melody 4, which also ends in a large implicative interval, the model only correlates significantly with adults ($r = .480$) and older children ($r = .582$). Interestingly, these correlations are robust to additional learning. With 30 more epochs of training, the correlations change an average of only .004.

The correlations suggest that the model settles quickly into a state that creates expectancies that are most like the older children (see Figure 2) and, in the current architecture, are robust to further learning[1]. However, it is also clear that the model has difficulty in generating a psychologically valid expectation when given a large implicative interval, where pitch reversal is at its height in terms of creating expectations. To further investigate these results, we performed a multiple regression analysis, including pitch proximity and pitch reversal as predictors of the model's expectancy ratings, as Schellenberg et al. did for step two of a hierarchical regression for the human ratings. The analysis revealed that the two-factor model is indeed a significant predictor of the model's rating behavior, $F(2, 59) = 49.45$, $p < .05$, adjusted $R^2 = .622$. However, only the proximity factor was a significant contributor to the model, t (59) = -9.90, $p < .05$. This regression confirmed that the model behaves according to the principle of pitch proximity, but not reversal.

Inspection of the pitch profiles generated by the model, compared to the profiles of older children provides visual indication of statistical findings (see Figure 2). While the expectancy ratings of notes within the same octave as the final tone (the center point of the graph) are high, there is a sharp drop off between octaves, while the ratings of the humans show smoother transitions in ratings between octaves. Further, expectancy ratings in the model do not show an elevation at the first note of the implicative interval (denoted by an asterisk on the x-axis) or surrounding tones, which is predicted by the principle of pitch reversal.

A second multiple regression analysis revealed that adding the simple frequency of a note in the training set as a predictor variable does improve the predictive power of the model. This three-factor model is again a significant predictor of rating behavior, $F(3, 59) = 35.92$, $p < .05$, adjusted $R^2 = .640$. Further, the simple frequency factor is a nearly significant contributor to the model, t (59) = 1.97, p = .05, while pitch proximity remains a significant contributor, t (59) = -6.58, $p < .05$. Thus the model was not simply tracking stimulus frequency, but it does contribute to its expectancies.

The simple correlations and multiple regressions indicate that the model is producing ratings that conform most to those that are predicted by Schellenberg's (1997) proximity principle and which are most like the profiles of older children, who have also been shown to produce profiles that are primarily predicted by the proximity principle. The regressions also revealed that the model's ratings are influenced by the melodic context of the test sequence, not just the simple frequencies of notes in the training set. We turn now to a more in-depth analysis of the training set to examine whether the data set differentially embodies pitch proximity, or if pitch reversal could be reasonably expected to emerge from this set.

Corpus Analysis

The main motivation for this analysis is to determine the prevalence of large implicative intervals compared to simple frequencies and small intervals, both of which have proven to be significant predictors of the model's expectancy ratings. These data are also interesting in and of themselves, as they reveal information about the degree of support for the theoretical principals being examined here in the intervallic statistics of real songs.

In terms of simple frequencies, the 4 most frequent notes (A_4, G_4, B_4, and D_5 respectively) each occur more than 1,000 times. The next most frequent are 11 notes that occur between 100 and 900 times. This second group includes only two sharps ($F\#_4$ and $A\#_4$) and all of the tones are in the fourth and fifth octave. The least frequent group includes most of the sharps in the corpus as members, and the only two notes that fall outside of the fourth and fifth octave (B_3 and A_3).

Moving to note intervals, 4,263 of the transitions are small intervals and the two most common intervals involving pitch movement are +/- 2 semitones (1,129 and 1,226 occurrences, respectively). However, even more frequent than these two interval sizes are unison intervals, which occur 1,657 times. A visual inspection of Figure 2 indicates that the model appears to favor movement over laterality.

Table 1: Correlations between the ratings of the model after one epoch of training and the three human groups. * p < .05.

Melody (Interval Size)	Adult	Old Child	Young Child
1 (Small)	.829*	.843*	.761*
2 (Small)	.699*	.866*	.180
3 (Large)	.515*	.602*	.188
4 (Large)	.480*	.582*	.134

[1] Support for this claim can also be gathered from a procedurally-identical simulation which was done with an SRN with 100 hidden units that was trained for 10,000 epochs at a learning rate of .001 and produced approximately the same results at test.

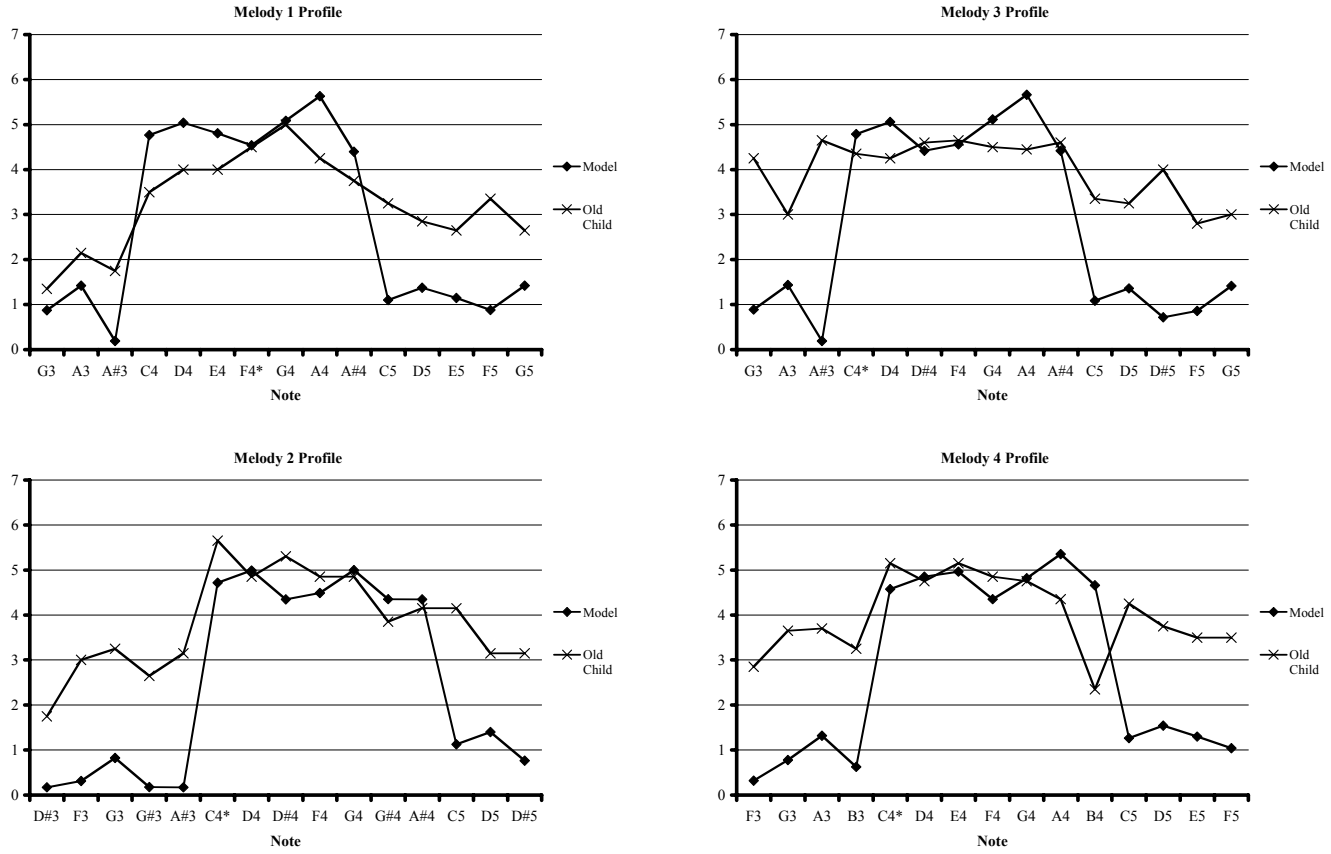

Figure 2: Rating profiles for the SRN and older children from Schellenberg et al (2002). The center note on the x-axis is the last note of the melodic fragment, the asterisk denotes the next-to-last note heard. The profiles of the small implicative intervals (Melody 1 and Melody 2) are on the left, the profiles for the large intervals (Melody 3 and 4) are on the right.

In terms of large intervals, there are only 3,255 large intervals, or slightly more than one-third of the corpus. Given a large implicative interval, the realization interval does tend to return back towards the first note of the implicative interval (there are 1,412 such returns), but it is also just as likely to stay near the last note. A small interval following a large interval occurs 1,240 times and is only slightly more likely to be a reversal of pitch contour. There are also 514 unison tones following a large interval, which means that proximity is a general property of the data set, rather than one that is specific to small or large intervals, and is thus learned by the model as this more general property, blocking the learning of the much lower frequency large returns following large intervals.

While large intervals are a large part of the data set, they are not as common as smaller intervals (including unisons). This presents the question of whether more sensitive listeners, who adhere to pitch reversal, have greater experience with large implicative intervals or if the responsible learning mechanisms are able to extract these less dominant characteristics of music.

Discussion

In the area of music cognition there are descriptions of perceptual principles that dictate human expectations for musical notes and which may be a result of learning. The present investigation determined the extent to which an SRN acts in accordance with such principles following training on a small set of musical sequences. This project was undertaken in the spirit of Schellenberg's (1997) simplification of a larger, more specific theory of musical expectancy as well as the spirit of the many cognitive scientists who have applied SRNs to procedural learning tasks.

The expectancy ratings generated by the model after just one epoch of training were most like those generated by older children in Schellenberg et al. (2002) and were indicative of a system operating according to a combination of simple frequency and pitch proximity, but not pitch reversal. An analysis of the training set revealed that pitch proximity indeed appears to be a general property of both large and small implicative intervals, rather than a principle which is specific to one or the other. It also revealed that pitch reversal is a much smaller component of the training set

than proximity, leading to the question of how the principle emerges in more experienced listeners.

Future work on this model should be directed towards providing more information to the network in order to establish how these factors might influence the expectations of the network. For instance, adding a self-organizing map that interacts with the output layer of the model may provide the model with top-down information regarding modes and keys, would allow the model to make contact with previously mentioned models, and would implement Narmour's (1990) top-down versus bottom-up distinction. Another line of inquiry may be in determining the extent to which varied and more complex musical exposure might account for the experience-based findings of Schellenberg et al. (2002).

Another potential extension of the model could be to explicitly represent larger amounts of context by adding additional layers to the recurrent portion of the model. This would force the model to represent longer time sequences and potentially allow it to learn the principle of pitch reversal from the few examples found in the current training set. An additional benefit to this architecture would be that the model might begin to memorize individual songs because it would have some basis for recognizing songs and this memorization might allow the model to develop, rather than maintaining the stable state it currently displays (cf. Bharucha & Todd (1989).

In general, the current application of an SRN to the musical domain is a novel contribution in this area, which may help to provide information about the role of experience and learning in musical perception and the processes involved in the acquisition of melodic expectation. It also proposes a candidate learning process for the acquisition of melodic expectancy, procedural learning, and the success of the model in this task provides evidence that pursuing the exploration of an empirical relationship between expectation and performance on a procedural task may be a fruitful line of experimental inquiry.

References

Bharucha, J.J. (1994). Tonality and expectation. In R. Aiello & J.A. Sloboda. (Eds). *Musical perceptions*, 213-239. London: Oxford University Press.

Bharucha, J.J. & Todd, P.M. (1989). Modeling the perception of tonal structure with neural nets. *Computer Music Journal, 13,* 44-53.

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

Chambers, J.C. (n.d.) *JC's ABC tune finder*. Retrieved December, 15, 2005 from http://trillian.mit.edu/~jc/music/abc/findtune.html.

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

Elman, J.L. (1989). *Representation and structure in connectionist models* (Tech. Rep. No. 8903), San Diego, California: University of California, San Diego, Center for Research in Language.

Krumhansl, C.L. (1995). Music psychology and music theory: Problems and prospects. *Music Theory Spectrum, 17,* 53-80.

Krumhansl, C.L., Louhivouri, J., Toivianen, P., Jarvinen, T. & Eerola, T. (1999). Melodic expectancy in Finnish folk hymns: Convergence of behavioral, statistical, and computational approaches. *Music Perception, 17,* 151-196.

Krumhansl, C.L. (2000). Rhythm and pitch in music cognition. *Psychological Bulletin, 126,* 159-179.

Mansfield, S. (2000). *How to interpret abc music notation.* Retrieved December 15, 2005 from http://www.lesession.co.uk/abc/abc_notation.htm.

Mitchell, D. (1968). *Every child's book of nursery songs.* New York: Corwn Publishers.

Mozer, M. (1991). Neural network music composition by prediction: Exploring the benefits of psychophysical constraints and multiscale processing. *Connection Science, 6,* 247-280.

Narmour, E. (1990). *The analysis and cognition of melodic complexity.* Chicago: University of Chicago Press.

Page, M.P.A. (1999). Modelling the perception of musical sequences with self-organizing neural networks. In Griffith, N. & Todd, P.M. (Eds). *Musical networks*, 175-198. Cambridge, MA: MIT Press.

Raph, J. (1964). *American song treasury: 100 favorites*. New York: Dover.

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.

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

Seeger, R.C. (1948). *American folk songs for children.* New York: Doubleday.

Seidenberg, M.S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science, 275,* 1599-1604.

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 & Psychophysics, 59,* 1069-1076.