In C. Davis, T. van Gelder & R. Wales (Eds.), *Cognitive Science in Australia, 2000: Proceedings of the Fifth Biennial Conference of the Australasian Cognitive Science Society.* Adelaide: Causal, 2000.

A PITCH IN TIME: AN ARTIFICIAL NEURAL NETWORK OF MELODIC EXPECTANCY

CATHERINE STEVENS

MARCS Auditory Laboratories University of Western Sydney, Australia
E-mail: kj.stevens@uws.edu.au

SUE BECKER AND LAUREL TRAINOR

Department of Psychology, McMaster University
1280 Main Street West, Hamilton, Ontario L8S 4K1 Canada
E-mail: ljt@mcmaster.ca becker@mcmaster.ca

The development of expectancies during the unfolding of auditory patterns in time is a recognized but poorly understood aspect of human cognition. This study investigates development of pitch-based expectancies in melody prediction. Two artificial neural networks tested the hypothesis that expectancies, such as pitch proximity and pitch reversal as proposed by Narmour [12], can be learned from exposure to a musical environment. A multi-layered back-propagation network and a perceptron both performed at better than chance level. The way in pitch relations are acquired and represented in networks is discussed together with implications for future experiments and network models of musical pitch development.

1 Introduction

Variation in pitch is an essential element of both speech and music. The way that musical pitch relations, such as interval, melodic contour, scale and key, are represented and interact in human memory warrants detailed analysis. Studies of music cognition suggest that an unfolding musical event gives rise to pitch and time-based expectancies [1,3]. The present study uses artificial neural networks to investigate the way in which melodic expectancies develop from exposure to a particular auditory environment. These computer models were constructed to examine the hypothesis that the perceptual organisation principles specified in a current theory of music cognition can be learned and represented in a neural network. The theory is the implication-realization (IR) model of Narmour [12,13,14]. Our models explore Narmour's claims that many expectancies in melody cognition can be explained, to a large degree, by structures that arise from local tone-to-tone transitions. Narmour proposes that the identified principles are the "given code" for perception of melodic patterns. As an alternative, we examine the extent to which artificial neural networks can acquire and develop such a code.

1.1 A Theory of Melodic Expectancy: The Implication-Realization Model

Narmour [12,14] provides comprehensive description and analysis of melodic structure in western tonal music. On the bottom-up side, he contends that melodies are perceived and cognized according to a number of universal principles that have their origin in Gestalt principles of organisation, such as proximity, similarity and closure. The principles apply to local, note-to-note transitions of melodies and characterise sets of possible continuations or implications suggested by an incomplete musical pattern. The bottom-up rules are relevant to all styles of melody whereas the top-down rules of the theory invoke specific stylistic structures. The current study concentrates on evidence for, and the explanatory power of, the bottom-up principles.

In the IR model, melody cognition is described as a series of closures, implications, and realizations. Closures are points of musical resolution; an implicative interval is one that is not closed and therefore sets up expectations for which note will follow; a realization is the actual note (or interval) that follows an implicative interval. Five principles reflect the main melodic implications of Narmour's theory [12]. However, listeners' expectations appear to be captured, in the main, by two principles derived from these five *pitch proximity* and *pitch reversal* [17,18]. *Pitch proximity* states that when listeners hear a small implicative interval in a melody they expect the next tone to be proximate in pitch to the second tone of the implicative interval (i.e. they expect a *small*-sized interval). *Pitch reversal* extends the pitch proximity principle to relations between non-adjacent tones and relates to expectancies that arise when intervals violate the pitch proximity principle, i.e. when a *large* implicative interval is heard. Once the coherence of the melody has been disrupted by a melodic leap, listeners expect a reversal of pitch direction. Therefore the *pitch reversal* principle proposes that listeners expect a change of direction *and* a relatively small interval.

Some experimental evidence has been gathered that suggests that listeners are sensitive to the IR principles [5,17,19]. If an artificial neural network is to be a valid model of human melody cognition then it too should display the principles. An artificial neural network as a statistical learning device provides a testable, mechanistic account of the way in which such principles develop. According to Narmour [12] the IR principles reflect a "genetic code" that operates when humans listen to melodies. The neural network approach provides an alternative account. It is possible to hard-wire a network and to examine performance of a network that contains a pre-established pattern of connectivity and weights. A more parsimonious method is to investigate the degree to which a network can learn such connectivity and learn to represent relevant information from mere exposure to a training environment. Both the network architecture and/or the nature of the training environment can be manipulated as independent variables in an effort to identify the connectivity and/or the experience required to develop particular sensitivities or behaviours. Specifically, we ask the question: what network architecture and what kinds of experience are necessary and sufficient for a network to display IR expectancies?

1.2 The Development of Melodic Expectancies in Two Artificial Neural Networks

It is hypothesized that if the artificial neural network is a valid model of human melody cognition then, after exposure to examples of Western tonal melodies, the network will construct a set of connection strengths and hidden unit activations that permit: a) prediction of the next note in familiar melodies; and b) prediction of the note following an implicative interval that conforms with Narmour's principles. Two different networks were constructed and exposed to identical sets of training and test patterns. The first network was designed to encourage the construction of an interval code where the pitch distance between each two successive tones was used and actual pitch values ignored. Interval is a higher-order feature that must be built up from simper features. For example, we imagine that a unit that detects an interval of a fifth would be built up by combining the outputs of hidden units in lower layers that each detect particular pitches that are a fifth apart in particular keys. The higher-level hidden unit can then recognize the interval of a fifth, regardless of key.

2 Method 1: Feed-forward 48-24-12-24 Back-Prop Network

2.1 Network Architecture

The multi-layered network consisted of four layers of units: an input layer containing 48 pitch units, two layers of hidden units containing 24 and 12 units respectively, and 24 output units. The task for the network was to predict the next note in a series of English folk melodies [9]. The input units of the feed-forward network stood for pitch classes of the Western tonal scale. To emulate the unfolding of music in time and, in keeping with Narmour's hypothesis that listeners process all melodic intervals as primitive, bottom-up generative events [13], the input units represented the occurrence of specific pitch classes at two successive timesteps. For each of the timesteps, T1 and T2, there were 24 input units covering two octaves of pitch classes from A3 to G#4. The input of a particular melody involved sequential activation of particular pitch classes at T1 and T2 and the network was trained to predict the next pitch class given the pitch classes active at T1 and T2. The 24 output units represented two octaves of pitch classes ranging from A3 to G#4.

Two layers of hidden units provided non-linearity to the system. The second layer of hidden units should provide a reconstruction of the input to enable correct prediction. Direct connections from T2 to the output units were added to facilitate this reconstruction by enhancing the immediate context that precedes the desired note. The feed-forward network was trained using the back-propagation of error learning algorithm with a learning rate of 0.02. The activation function was continuous sigmoidal. Low-level perceptual processes precede interval prediction and there are existing neurophysiological and artificial neural network models that specify the mechanism and neural organisation required to achieve the task [2,7,11,16]. Such perceptual mechanisms are assumed to provide the front-end to the models presented here.

2.2 Training Patterns

The training patterns consisted of a quasi-random selection of folk melodies from Sharp's collection of English folk songs [9]. The 15 short melodies conformed to Western tonality, ranged from approximately 20 to 50 notes in length, and included two examples from the major scales A, C, D, E, F, G, E-flat, and one example of a melody in B-flat major. As one aim of the study was to investigate the degree to which networks can represent pitch relations such as scale or key, the training melodies retained their original key rather than being transposed into a single key. All note durations in the melodies were equal and, as we were interested in the prediction of intervals, note repetitions were also omitted. The training set consisted of 608 pitch events. Given the melody statistics, of the 608 training patterns, 566 or 93% could potentially involve proximity and 36 events or 6% of the training patterns could have involved reversals. The actual realisation of these principles in the training set melodies was 409/566 or 72.3% of possible proximity relations, and 24/36 or 66.6% of possible reversals.

After a series of training cycles, the network was tested in two ways: accuracy of prediction of the folk melodies was assessed and the network was exposed to all possible musical intervals and prediction of the next note was recorded. There were 264 intervals representing intervals from 1 to 11 for each of the 12 pitch classes in both ascending and descending order.

3 Results: Model #1

After 15,000 training cycles the network recorded a total summed squared error of 343.67 with gradient of 8.52. Performance of the network was measured by taking the "best guess" or the output unit with the highest activation at each timestep. The accuracy of the best guess prediction of the next pitch event in the folk melody training set was 54.8%. Performance at chance level would be approximately 1/12 or 8%. The 54.8% level of performance was regarded as satisfactory: it was undesirable for the network to fully memorise the training set and it would be unlikely that perfect prediction would be achieved normally by listeners. A prediction was made by the network for all input patterns. Although prediction was correct, on average, for one of every two events, an additional 6% of incorrect predictions involved activation of a neighbouring pitch class. Another 6% or so of errors included activation of the correct pitch class together with competing activations from other pitch classes.

3.1 Analysis Using Narmour's Principles

Of specific interest was the extent to which a network exposed to exemplars of Western tonal music would show evidence of the principles of melodic cognition discussed by Narmour [12,13] and Schellenberg [17,18]. The following analysis of network performance was based on Schellenberg's [18] two-factor account of the implication-realization model. The trained network was exposed to all possible ascending and descending musical intervals and evidence for the prediction of intervals according to proximity and pitch reversal IR principles was scored. Proximity was scored for those intervals up to and including 5 semitones: a score of 1 (correct) was given where an interval within this band was followed by an interval of up to 3 semitones in either direction. Pitch reversal was scored for those intervals of seven semitones or more: a score of 1 (correct) was given where an interval that followed a large interval reversed direction and was smaller in magnitude than the one that preceded it. The interval test set consisted of 120 intervals that implied movement by proximity and 119 intervals that implied pitch reversal. Table 1 shows scores for proximity and pitch reversal.

Table 1. Back-prop model prediction of pitch proximity and pitch reversal

Principle	Raw Score	Percentage	Chance
Pitch proximity	68/120	56.7%	25%
Pitch reversal	78/119	65.5%	25%

Given the relatively low incidence of implied and realized pitch reversals in the training set the above results indicate that the back-prop network is performing well in predicting pitch reversals. These "exception" cases seem to have been effectively learned in the multi-layer network. Prediction of intervals according to proximity was also better than chance.

4 Method 2: Single-Layer Perceptron

For comparison, it was relevant to examine the degree to which performance improves or declines using a simple, single-layer linear perceptron. The multi-layered network may have solved the pattern classification problem by forming an analog code for pitch in the hidden layer. If an analog representation is used it may be possible for such information to be represented in the weight matrix of a single-layer network. A perceptron was constructed and exposed to an identical set of melody training and interval test patterns.

4.1 Perceptron Architecture

The single-layer network consisted of 48 input and 24 output units. The input units stood for the pitch classes that spanned two octaves from A3 to G#4, for Time 1 and Time 2. The output units represented 24 pitch classes from A3 to G#4 and the task for the network was to again predict the next pitch class in a series of 15 folk melodies. The learning rate of the perceptron was 0.2 and, after just 4,100 training cycles, the prediction accuracy of the network based on the best guess was 51.2%.

5 Results of the Perceptron Model

5.1 Prediction and IR Principle Results

The perceptron was also tested using the interval test patterns and the output unit with the highest activation was examined for its conformity with Schellenberg's proximity and pitch reversal principles. Results are shown in Table 2.

Table 2. Perceptron prediction of pitch proximity and pitch reversal

Principle	Raw Score	Percentage	Chance
Pitch proximity	77/120	64.2%	25%
Pitch reversal	84/119	70.6%	25%

The perceptron was able to predict the next pitch event in a melody as well as the back-prop network. Similarly, output in response to particular musical intervals conforms well with the principles of melodic expectancy discussed by Narmour and validated by Schellenberg. For both the proximity and the reversal principles the linear network outperforms the back-prop network. There was also a tendency for both networks to respond most effectively to intervals that imply pitch reversal.

5.2 Perceptron Weight Analysis

One of the advantages of the perceptron is that it is possible to scrutinise the connection strengths from all input units to each output unit to determine the way in which the task is being achieved. In most instances, and not surprisingly, the connections from the final 24 input units (T2) mirror the connection strengths from the first 24 input units, as all pitch events are represented initially at T1 and then translated to T2. There are occasions where the T2 weights differ slightly from T1 and this may reflect the way in which the perceptron is able to code pitch events as they stand in local temporal relations to one another. Even with a narrow context window of two events some temporal relations appear to be represented. The addition of greater context at input and/or output would further enhance the representation of higher-order temporal relations and, ultimately, accuracy of prediction performance.

In terms of musical keys, the weights that have developed for each output unit reflect the hierarchy of importance of tones relative to a particular output tone or tonal center [10]. For example, weights from inputs to output unit #6 (D) include positive weights for tonally important pitch classes in the key of D major: the tonic D, the mediant F#, and leading note C# (Figure 1). Weights that link input units to output unit A are all negative except for input units that stand for pitch class A. The tonic of C major is heavily weighted in the matrix from input units to the C output unit. The weight matrix for output unit #3, B is surprising with the development of positive weights to members of the key signature F#, C# and D# despite no melody in B major being included in the training set! Weights to distinctive or stable tones are excitatory and those that are *not* characteristic of a particular key appear to be inhibitory. Throughout the course of training, the tones that are stable and tonally important remain with an excitatory value.

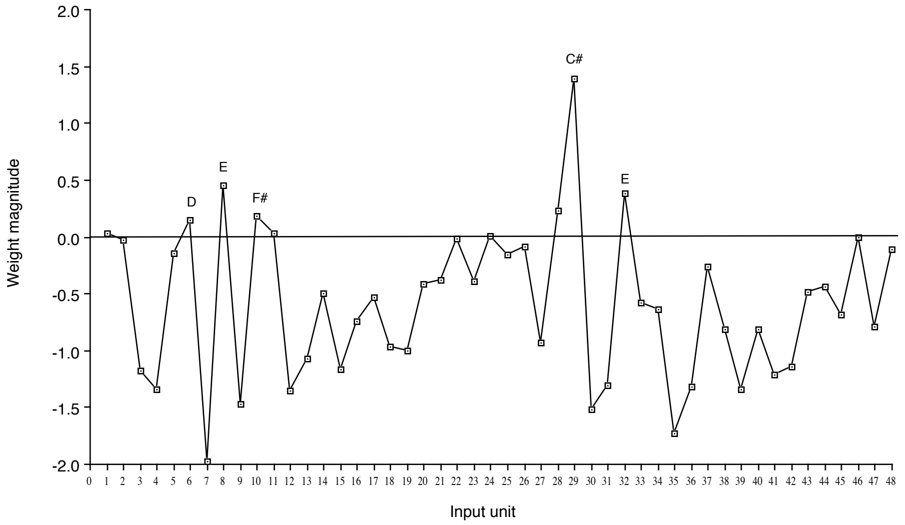

Figure 1. Connection strengths from 48 input units to output unit for pitch class D.

5.3 Comparison of the Models and Two Tests of Generalisation

The back-prop and linear networks were tested on two generalisation tasks. The first was a test of note prediction using a melody from the training set that had been transposed from F major to the closely-related dominant key of C major (melody no. 265B; original accuracy 14/26 correct prediction). The linear and back-prop networks showed comparable prediction performance: 29.2% for the linear network and 26% prediction accuracy for the back-prop net. Prediction was poorer than for the original melody but still better than chance (8%). Given the generally poorer recognition we can assume that neither of these networks represents much of the higher-order structural relations within the melodies and there is little representation of relative pitch and key information independent of absolute pitch. A more encouraging result is that the best guess output activations of each network conform well with the melodic expectancy principles of proximity and reversal. In instances where proximal notes would be expected, the linear net produced 13/18 (72%) proximal or symmetrical responses and the back-prop net produced 12/18 (66%) proximal responses. Pitch reversals were implied four times in the melody and the linear network filled the gap with a reversed and smaller interval three out of four times (75%), while the back-prop network realised the pitch reversal on four out of four occasions (100%).

Although the networks had been tested on discrete intervals they were also exposed to a new melody that had not been included in the training set. This short fragment of a folk melody [17, p.84, fragment #1] contained 11 notes and was in the key of D major. Clearly, it would be impossible for a network or human to predict with perfect accuracy a previously unheard melody but we were interested in the nature of the predictions and the adequacy of the network representation for such a task. The linear network scored 5/11 correct note predictions (45.5%) and the highest output activations were consonant with the key of D major. In similar fashion, the back-prop network yielded a prediction accuracy of 44.4% and, again, the most active output units reflected the key signature of D major. Performance on this task will be a product of the degree to which the new melody is predictable, tonally stable, and the extent to which the network has an adequate representation for the particular key. Based on this single example, both networks appear to have developed weights that permit good general prediction of tonality and generate predictions that conform with the principles of proximity and pitch reversal from 44% (for novel melodies) to 54% (for familiar melodies). It would be useful to compare network performance with predictions or ratings obtained from musically trained and untrained listeners assigned a comparable task.

6 General Discussion

The neural network models reported here are an existence proof that the principles central to Narmour's model of bottom-up melodic expectancy can be learned by exposure to, and representation of statistical regularities within, a subset of Western tonal melodies. Some evidence has been provided of the sensitivity of relatively simple neural networks to pitch proximity and pitch reversal. Although Narmour acknowledges the importance of both bottom-up and top-down information in melodic processing, the present results illustrate the parsimony and explanatory power of low-order, note-to-note transitions. Narmour [13] has referred to the IR principles as a genetic code whereas the neural network approach demonstrates one way that melodic expectancy may develop through exposure to a musical environment. The networks, like Narmour's theory, reflect the statistical regularities of pitch structure in Western tonal music. The structure of the music itself, reflecting as it does human perceptual and cognitive limits and constraints, is likely to be one of the best indicators of a "genetic" code.

With even a simple, linear network it has been possible to represent some of the regularities and structure inherent in a random selection of musical compositions and predict the next note in these melodies at better than chance level. This corroborates Gluck & Bower's [6] plea for parsimony in the design of artificial neural networks. The network is easily interpreted and may provide an adequate mechanism for the extraction and representation of regularities of one component or aspect of the auditory process. A more computationally powerful network will be required when additional dimensions such as duration and amplitude are input, and a linear system is unlikely to be able to recognize familiar melodies that have been transposed along the pitch dimension. The complete analysis of the perceptron afforded by its simple structure also satisfies Coltheart's [4] claim that it is the "functional architecture" of a connectionist model that is of psychological relevance and warrants full analysis.

Given the prevalence of notes moving by step in the training set, network results indicate that pitch reversal "exceptions" were learned effectively. The existence of exceptions in music, such as non-stepwise movement and unexpected modulations, motivates design of a model of melodic expectancy consisting of adaptive mixtures of local experts [8]. Machine learning algorithms such as CLARET, developed for spatio-temporal sequences, also have potential for modelling complex pitch relations [15]. A further refinement to the present model would be to build proximal pitch relations into the network by coarse-coding pitch input units.

The important next step is to develop simulations of experimental tasks, such as those conducted by Carlsen [3,20], and compare network output and human performance. Although there is an abundance of experimental studies on pitch perception, interval and contour recognition, key classification, less is known about the relations between pitch, interval, scale step and key and the way these relations are learned and represented in memory. Artificial neural networks have much to offer as testable, predictive models of possible mechanisms that underpin acquisition of pitch relations and interval structure, and as tools to explore the differential effects of age and experience.

7 Acknowledgements

This research was supported by a UWS Macarthur Research Encouragement Award.

References

1. Bharucha J. J., Music cognition and perceptual facilitation: A connectionist framework, *Music Perception* **5** (1987) pp. 1–30.
2. Bharucha J. J. and Mencl W. E., Two issues in auditory cognition: Self–organization of octave categories and pitch–invariant pattern recognition, *Psychological Science 7* (1996) pp. 142–149.
3. Carlsen J. C., Some factors which influence melodic expectancy, *Psychomusicology* **1** (1981) pp. 12–29.
4. Coltheart M., Connectionist modelling and cognitive psychology, http://www.cs.indiana.edu/Noetica/OpenForumIssue1/Coltheart.html (1995).
5. Cuddy L. L. and Lunney C. A., Expectancies generated by melodic intervals: Perceptual judgments of melodic continuity*, Perception & Psychophysics* **57** (1995) pp. 451–462.
6. Gluck M. A. and Bower G. M., Evaluating an adaptive network model of human learning, *Journal of Memory and Language* **27** (1988) pp. 166–195.
7. Griffith, N. Development of tonal centres and abstract pitch as categorizations of pitch use. In N. Griffith and P. M. Todd (eds.), *Musical networks: Parallel distributed processing and performance* (MIT Press, Cambridge, Mass., 1999) pp. 23-43.
8. Jacobs R. A., Jordan M. I., Nowlan S. J. and Hinton, G. E., Adaptive mixtures of local experts, *Neural Computation* **3** (1991) pp. 79–87.
9. Karpeles M., *Cecil Sharp's collection of English folk songs Vols 1 & 2* (Oxford University Press, London, 1974).
10. Krumhansl C. L., The psychological representation of musical pitch in a tonal context, *Cognitive Psychology* **11** (1979) pp. 346–374.
11. Moore B. C. J., *An introduction to the psychology of hearing* (4th ed., Academic Press, San Diego, 1997).
12. Narmour E., *The analysis and cognition of basic melodic structures* (University of Chicago Press, Chicago, 1990).
13. Narmour E., The melodic structures of music and speech: Applications and dimensions of the implication–realization model. In J. Sundberg, L. Nord, and R. Carlson (eds.), *Music, language, speech and brain* (Macmillan, 1991) pp. 48–56.
14. Narmour E., *The analysis and cognition of melodic complexity: The implication–realization model* (University of Chicago Press, Chicago, 1992).
15. Pearce A. R., Caelli, T. and Goss, S., On learning spatio-temporal relational structures in two different domains. In R. Chin and T.-C. Pong (eds.), *Computer vision-ACCV'98: Lecture notes in computer science Vol. 1352* (Springer Verlag, 1998) pp. 551-558.
16. Sano H. and Jenkins B. K., A neural network model for pitch perception. In P. M. Todd and D. G. Loy (eds) *Music and connectionism* (MIT Press, Cambridge, Mass., 1991) pp. 42-49.

17. Schellenberg E. G., Expectancy in melody: Tests of the implication–realization model, *Cognition* **58** (1996) pp. 75–125.
18. Schellenberg E. G. Simplifying the implication–realization model of melodic expectancy, *Music Perception* **14** (1997) pp. 295–318.
19. Thompson W. F., Cuddy, L. L. and Plaus, C., Expectancies generated by melodic intervals – Evaluations of principles of melodic implication in a melody completion task, *Perception & Psychophysics* **59** (1997) pp. 1069–1076.
20. Unyk A. M. and Carlsen J. C., The influence of expectancy on melodic perception, *Psychomusicology* **7** (1987) pp. 3–23.