A Connectionist Model of Intermediate Representations for Musical Structure

Edward W. Large
Computer and Information Sciences Department
2036 Neil Avenue
large@cis.ohio-state.edu

Caroline Palmer
Psychology Department
1885 Neil Avenue
cpalmer@magnus.acs.ohio-state.edu

Jordan Pollack
Computer and Information Sciences Department
2036 Neil Avenue
pollack@cis.ohio-state.edu

The Ohio State University
Columbus, OH, 43210

Abstract

The communication of musical thoughts and emotions requires that some musical knowledge is shared by composers, performers, and listeners. Computational models of musical knowledge attempt to specify the intermediate representations required to generate adequate predictions of musical behavior. We describe a connectionist model that encodes the rhythmic organization and pitch contents of simple melodies. As the network learns to encode melodies, structurally more important events tend to dominate less important events, as described by reductionist theories of music (Lerdahl & Jackendoff, 1983; Schenker, 1979). We describe an empirical study in which improvisations on a tune by a skilled music performer are compared with the encodings produced by the network. The two are examined in terms of the relative importance of the musical structure they posit at intermediate levels of representation.

Introduction

A primary goal of music cognition is to specify mental representations for musical knowledge. Computational models of music composition, performance, and perception often posit multiple levels of structural description in mental representations. We refer to these levels of structural description as *intermediate representations*, because they mediate between the perception of a musical surface (the score or acoustic signal) and the resulting musical behavior (the performance or memory of a piece). However, the proposed structural descriptions often fall short of adequately specifying the relative importance of the musical events. We propose a connectionist model of mental representations for music that emphasizes the hierarchical nature of musical structure, and we compare its predictions of relative importance with evidence from skilled music performance.

Theoretical accounts of mental representations for musical structure often emphasize the importance of hierarchical organization. Hierarchical models of rhythmic organization, for example, describe the way in which musical events are combined to form larger structural units in a nested fashion (Cooper & Meyer, 1960). In a particular musical context, certain pitch events are heard as being dominant in the hierarchy and others are heard as elaborations of the dominant events (Lerdahl & Jackendoff, 1983; Schenker, 1979). Some perceptual cues to hierarchical organization are present in performed music, such as expressive variations in timing and dynamics. However, the models described above posit intermediate levels of mental representation that are based on information not necessarily present in the musical input. These intermediate levels of description are thought to reflect statistical regularities, derived from the input with the aid of general knowledge about the roles that events play in a particular musical idiom (Palmer & Krumhansl, 1990; Knopoff & Hutchinson, 1978).

Computational models of music cognition attempt to specify the intermediate representations required to generate adequate predictions of human behavior. Several researchers have adopted connectionist models which provide general-purpose learning algorithms capable of responding to the statistical regularities of the learning environment. However, connectionist models have been notoriously weak at representing hierarchical relationships, such as those found in music or language (Fodor & Pylyshyn 1988). Recursive Auto-Associative Memory (RAAM) is a connectionist architecture which develops distributed representations of hierarchical structures, directly attacking this problem of representational adequacy (Pollack, 1988; Pollack, 1990).

In this paper, we describe a RAAM model that encodes the rhythmic organization and pitch contents of simple melodies. As the network learns to encode melodies, structurally more

important events tend to dominate less important events, as described by reductionist theories of music (Lerdahl & Jackendoff, 1983; Schenker, 1979). We then describe an empirical study in which improvisations on a tune by a skilled music performer are compared with the encodings produced by the RAAM network. The two are examined in terms of the relative importance of the musical structure they posit at intermediate levels of representation.

Time-Span Reductions

One theory emphasizing intermediate levels of representation attempts to model an experienced listener's intuitions of Western tonal music (Lerdahl & Jackendoff, 1983). The theory describes many types of hierarchical representations, including metrical structure, grouping structure, and time-span reduction. Metrical structure describes the way in which a series of pulses are mentally combined to create nested hierarchical levels of alternating strong and weak pulses. Grouping structure describes nested groups of events forming motives, phrases, and larger sections of music. The outputs of metrical and grouping structures combine to segment a piece into hierarchically nested rhythmic units called time-spans. At the lowest levels time-spans are determined primarily by metric structure, and at the highest levels by grouping structure.

A *time-span reduction* organizes all musical events in a piece into a single coherent structure that reflects a strict hierarchy of relative importance. Within each time-span a single most important event is identified and all other events are heard as subordinate to it. In Figure 1, we show a time-span reduction for the melody "Hush Little Baby". The top staff shows the melody, and the brackets show how the piece is segmented into time-spans. The staves below show the intermediate levels of the reduction. At each level, less important events are eliminated, leaving a "skeleton" of the melody.

In the next section we propose a connectionist model for encoding representations of hierarchically nested time-spans. By examining the representations, we can predict the relative importance of musical events within each time-span. We then describe evidence from a skilled music performance that tests the predictions of relative importance made by the connectionist model. A pianist's improvisations on a theme are contrasted with the model's predictions of relative importance of different musical events.

Recursive Auto-Associative Memory

RAAM is a connectionist architecture that develops distributed representations of variable sized, compositional data structures. It has been used to model the encoding of hierarchical structures such as those found in linguistic syntax and logical expressions (Pollack, 1990). Conceptually, a RAAM consists of two machines, a compressor and a reconstructor. The compressor is trained to recursively encode sets of fixed-width patterns into single patterns of the same size. The reconstructor is trained to recursively decode the patterns produced by the compressor into facsimiles of the original sets of patterns. These mechanisms are co-evolved by linking their training sets together using an auto-associative form of back-propagation. Our current work is based on an implementation of a RAAM as a 3-layer feed forward neural network where the input-to-hidden layer transformation is the compressor and hidden-to-output layer transformation is the reconstructor.

In order to find the intermediate distributed representations for a set of melodies, we segment the melodies into time-spans as shown in Figure 1, and use these hierarchies as the training set to a RAAM. The primitive events in each melody are represented as binary feature vectors. We have chosen the encoding strategy shown in Figure 2. One set of units encodes pitch class, and a second set encodes local implied harmony

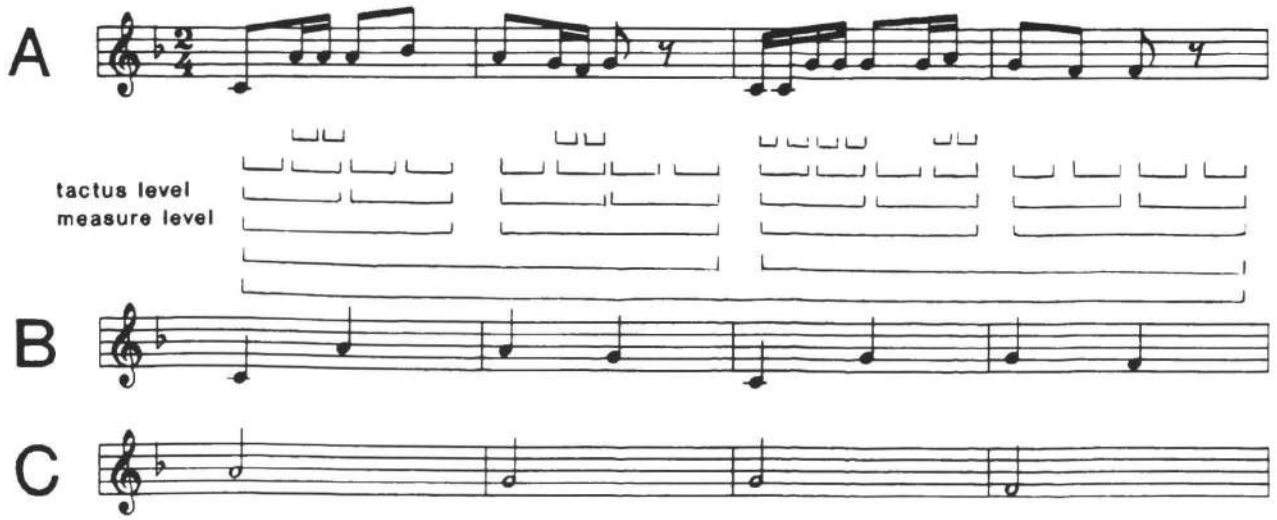

Figure 1: A Partial Time-Span Reduction for "Hush Little Baby."
A) Original melody; B) Tactus level reduction; C) Measure level reduction.

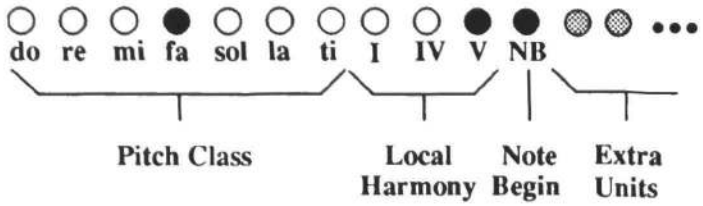

Figure 2: The representation of pitch events.

similar to Lerdahl and Jackendoff's use of local harmony in creating time-span reductions. An additional unit designates the beginning of an event. When this unit is turned on, it indicates that the event has its attack at that particular point. When turned off, it indicates that event is a continuation of a previous event.

After training, the compressor and reconstructor are treated as separate networks for the processes of encoding and decoding intermediate representations. Figure 3 shows a a short melodic excerpt segmented into time-spans, and depicts the processes of encoding and decoding it. First, the compressor encodes a set of the lowest level time-spans, and these encodings are recursively encoded to produce a representation for the entire structure. Next, the reconstructor decodes the compressed representation to retrieve a facsimile of the original structure. In order to capture the distinctions between binary and ternary groups found in music, we use a quaternary RAAM, that is, one with four fields of input units (11 input units per field) activated as shown in Figure 4.

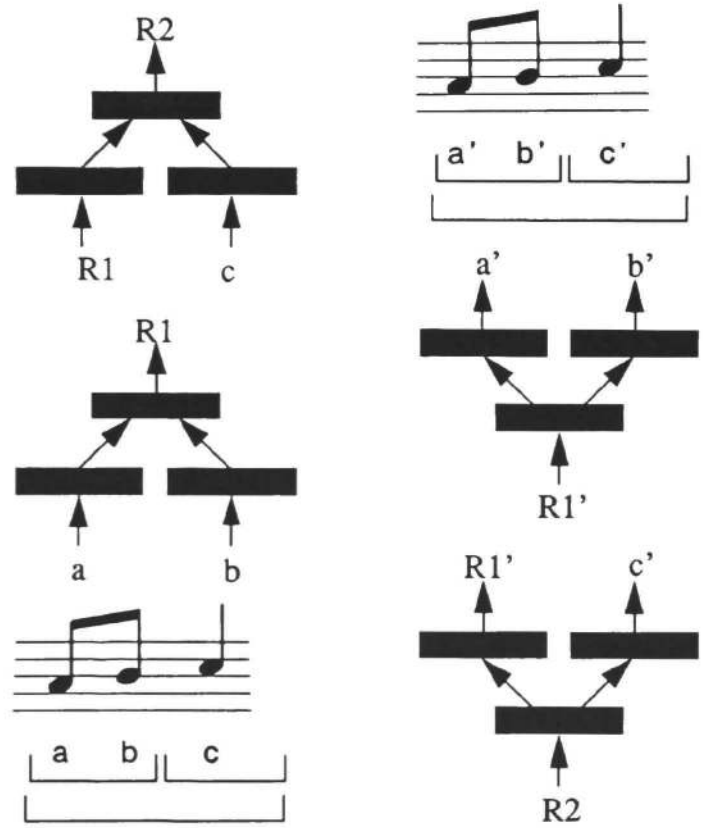

Figure 3: Network encoding and decoding of a time-span

The time-span [[a b] c] is encoded by compressing the primitive event representations of a and b, producing representation R1. R1 is then compressed with the next event, c, producing R2, which is a representation of the higher level structure. Next, R2 is decoded by reconstructing facsimiles R1' and c'. R1' is then reconstructed to produce facsimiles a' and b'. Thus a facsimile of the original structure, [[a' b'] c'], is produced.

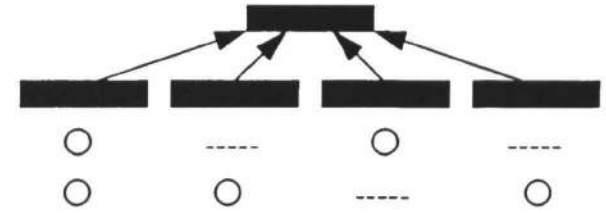

Figure 4: Encoding both binary and ternary groups in a quaternary RAAM

When a binary group is encoded (A), the first and third fields are activated. The second and fourth are set to zero, simulating a binary RAAM. When a ternary group is encoded (B), the first, second and fourth fields are activated. The third is set to zero, simulating a ternary RAAM.

The encodings developed by the network reflect the temporal embedding of the time-span hierarchy. The duration of each encoding produced by the network is considered to be the sum of the durations of its component events. This allows us to represent the rhythmic structure of the melody without pre-specifying a smallest possible time-slice (see Todd, 1989). Instead, we follow Lerdahl and Jackendoff's description of the tactus as the most salient metrical level (i.e. the level of foot-tapping tempo). We require that the tactus be continually represented throughout the piece, but time-spans derived from smaller metrical levels are represented only when actually present in the melody.

The compressor and reconstructor networks, taken together, comprise a well-formedness test for novel structures. Given a novel structure, the compressor network is used to create a representation. The reconstructor network is then applied to the representation to retrieve its constituents. If the reconstructed structure matches the input structure, either exactly or within some tolerance, this novel structure can be considered to be well-formed. We will use the difference between the constituents of the input structure and those of its reconstruction to determine the relative weighting of each musical event within the representations developed by the network.

Measures of Relative Importance

RAAM Network

Sixteen nursery tunes (such as "Mary Had a Little Lamb") were chosen as a training set because they provide a simple, natural musical case for study. Each tune was a simple melody between 4 and 12 measures in length, with a meter of 2/4, 4/4, 6/8, or 12/8. The tunes comprised ten unique melodies;

four of these ten melodies had variations (tunes with similar pitch or rhythmic properties). Although the event representations required only 11 bits, we used 23 units, allowing 12 extra "degrees of freedom" for the system to use in arranging its intermediate representations. These extra dimensions of representation were set to 0.5 on input, and trained as don't-care's (Jordan 1986) on output. Thus, for the quaternary RAAM, our network had 92 input and output units, and 23 hidden units. The network was trained on the time-spans for the melodies with length less than or equal to the measure level. The network was not trained until it memorized the tunes, but instead for 1500 cycles with a learning rate of 1.0 and momentum of 0.5. The network was therefore not able to reconstruct every melody in the training set note-for-note. The network's output representations are interpreted from the output values of the pitch-class units produced by the decoder for each event in the sequence. If all outputs are less than some threshold value at any given point, the event is interpreted as a rest (a null event). Otherwise the pitch class unit with the greatest activation at any given point is interpreted as the reconstructed event. With this interpretation the reconstructor was able to reproduce fairly accurate facsimiles of input melodies, although some events were "forgotten" and others were "remembered" incorrectly (in the reconstruction but not from the original melody).

To test the network's ability to encode novel tunes, we exposed it after training to a melody dissimilar to the original set of 14 tunes. Figure 5A shows the melody "Hush Little Baby" after its reconstruction by the network, as described above. Because this tune was dissimilar to those in the training set, we used the sensitive threshold activation value of 0.1. The network's actual activation values for each event in the sequence are shown below the reconstruction.

Improvised Performances

To compare the network's predictions of relative importance with those of skilled musicians, we recorded improvisations on a tune by a skilled pianist. Improvisation in Western tonal music commonly requires a performer to identify a framework of important melodic and harmonic events, and apply procedures to create elaborations and variants on them (see Pressing, 1988 for a review of improvisational models). Thus, improvisation on a musical tune allows the pianist freedom to determine which musical events should be retained (those of primary importance), and which should not (those of less importance).

A professional pianist from the Columbus, Ohio area was asked to perform three nursery tunes on a computer-monitored acoustic upright piano. The pianist was experienced in improvising in a contemporary/popular musical style. The pianist performed three melodies: one included in the network's training set, one not in the training set but similar to it in pitch and timing (a variation on one of the melodies), and a third melody unrelated to any in the training set. The pianist first performed each melody as it was notated, to become familiar with it. He then improvised five simple melodic (single-line) variations on each melody. All pitch, intensity, and timing information in the performances was recorded on computer, and compared with the original melody. Only performances of the third melody (the most stringent test of the network's ability to predict relative importance) will be described here.

According to our application of the time-span reduction hypothesis to improvisation, structurally important events should be less likely than unimportant events to change in variations of a melody. To test this hypothesis, the number of musical events identical to the original melody in pitch were summed across variations. The sums ranged from 5 (the same

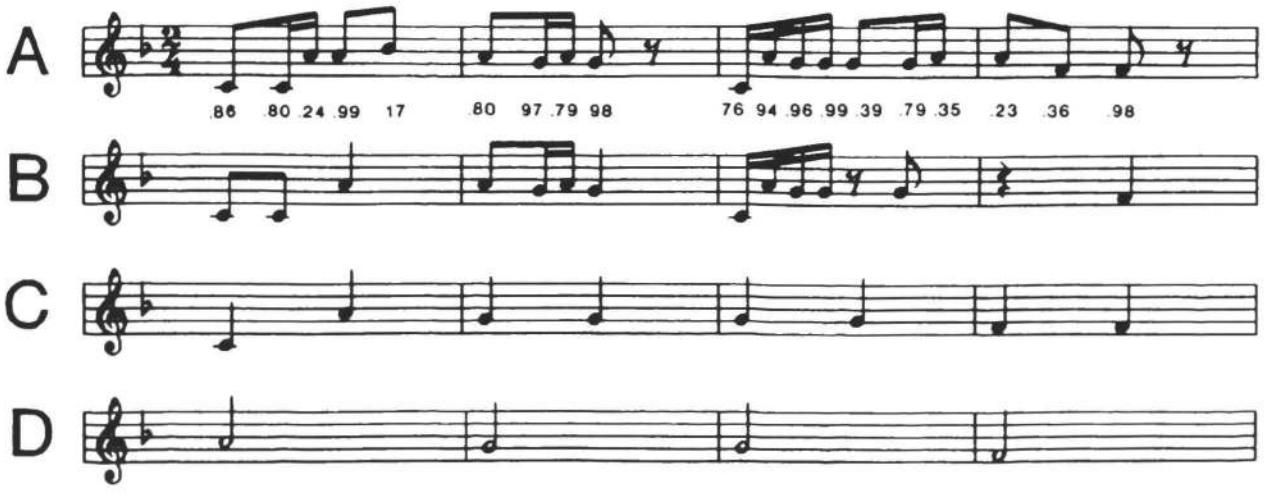

Figure 5: "Hush Little Baby."
A) Network reconstruction; B) 0.66 threshold criterion reduction;
C) Lerdahl & Jackendoff style tactus level reduction; D) Lerdahl & Jackendoff style measure level reduction.

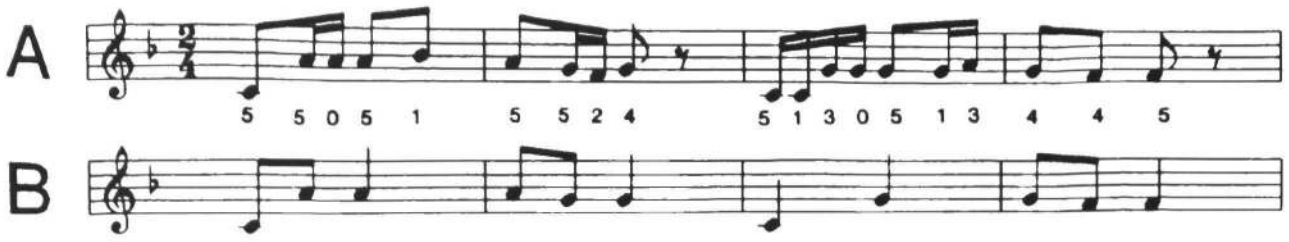

Figure 6: "Hush Little Baby."
A) Original melody with performance ratings; B) Rating criterion reduction.

pitch in each of the five performances) to 0 (the same in no performance) for each event location. Figure 6 shows the original melody, with the ratings for each location.

Comparison of RAAM Network and Improvised Performances

First, we compare the original tune (Figure 6A) and the network's reconstruction (Figure 5A). The reconstructed tune is a reasonable facsimile, considering that it was not a variation of any tune from the training set, but a distinctly novel melody. This indicates the network's ability to generalize beyond its input. We then compare the network's predictions of relative importance with the pianist's improvisations by developing reductions for both the reconstructed melody and the improvisations. An intermediate representation was developed for the improvisations by including only events from the original melody that were retained often across performances (those scoring 4 or higher). Using this method, we included only the highest third of the range of values, and we applied the same criterion to the network output by raising the activation threshold to 0.66. The reductions obtained from this criterion are shown for the network in Figure 5B and for the improvisation data in Figure 6B. At this level the reductions show significant agreement. In general, the network tended to retain more events than the improvisations.

Finally, we can compare more abstract levels of representation to those predicted by Lerdahl and Jackendoff's theory. It is difficult to produce further reductions for the improvisations because of the resolution of our measurements. However, we can produce further reductions of the network's encodings. Instead of reapplying the threshold criterion, we chose the event in each time-span with the highest activation, similar to Lerdahl and Jackendoff's (1983) method of computing reductions. By comparing Figures 5D and C to Figures 1B and C, we see strong agreement at intermediate levels of reduction. The network's ability to generalize well enough to develop a representation for this melody may be related to its weighting of important events at each representational level.

Conclusions

The similarities seen here between improvisational music performance and a connectionist model of simple melodies may result from similar computational constraints. The skill of improvising on a theme has been described as a largely unconscious process of identifying important structural elements and applying creative procedures to elaborate on those elements (Johnson-Laird, in press; Steedman, 1982). The resulting variations are related to each other by transformational rules that generate the possible improvisations on that theme. One important consequence of this approach is that it reduces the memory demands that can accompany the use of multiple intermediate representations. Instead of retaining each element at each representational level (thereby increasing the necessary storage capacity), only a reduced set of elements is stored, from which other representational levels are generated. The RAAM network produces a compressed representation, in which the structurally more important events dominate, such that they are more likely to be reliably reconstructed than structurally less important events. These similar constraints on processing demands may account for the similarities seen here in the improvised variations in music performance and in the network reconstructions.

Reductionist theories of music cognition have inspired other computational models of intermediate representations. Scarborough et. al. (1989) describe a parallel constraint satisfaction approach for the perception of metric structure, modelled as the response of independent metronome-like agents to individual musical events. Based on inter-agent constraints that enforce Lerdahl & Jackendoff's rules for metric structure, a hierarchical representation of the metric structure of a piece emerges. Rosenthal (1989) has described the perception of grouping structure as the process of constructing "recognizer agents". The processes which construct recognizers operate in accordance with Lerdahl and Jackendoff's rules for producing grouping structure analyses. Once constructed, agents recognize the repetition of rhythmic structures, thus implementing a restricted form of parallel structure recognition. The program's output is a mental representation of the piece stored as symbolic data structures. However, additional mechanisms must be posited to determine the similarity of two elements in the model (Rosenthal, 1989). One of the advantages of using RAAM is that the intermediate representations admit simple similarity measures, such as euclidean distance, capturing the statistics of the input environment.

Other researchers in music cognition have focused on the importance of expectation. Meyer (1956) argues that "...the inhibition of the tendency to respond or, on the conscious level, the frustration of expectation (is) the basis of the affective and intellectual response to music." Bharucha and Todd (1989) have described a computational model of musical expectation using Jordan nets (Jordan, 1986) and Todd (1989) has described how similar networks may be used for music composition. Todd notes, however, that these networks produce melodies "...high in local structure, but lacking in overall global organization". These computational approaches to expectation may fail to capture the intermediate representations required for hierarchically structured events such as music.

Another observation we would like to make regards the adequacy of the RAAM architecture for developing representations of the hierarchical structure of melodies. Although we have only reported the model's results for the encoding of one tune, this result represents the most stringent test of the network's ability: to encode novel sequences in a manner similar to that of skilled musicians. We did not expect that this tune would be correctly encoded by the network because it was not closely related to the melodies in the training set. The fact that the network was able to reconstruct a reasonable facsimile of this melody shows that not only is the network capable of encoding melodic structure, but it is capable of generalizing in a robust manner. The final observation regards the nature of the representations that the network develops. The fact that the representations weighted events in a way similar to both the musician's choices of events to retain in improvisations and predictions of relative importance based on Lerdahl and Jackendoff's theory (1983) indicates that the computational model may capture the relevant hierarchical properties of humans' mental representations for musical melodies.

Acknowledgments

The research was partially supported by NIMH grant 1R29-45764 to the second author and ONR Grant N00014-89-J1200 to the third author. Thanks to Carolyn Drake and Mari Jones for their comments on an earlier draft of this paper, and to Peter Angeline, Kory Klein, and Carla van de Sande, for their assistance.

References

Bharucha, J. J. & Todd, P. M. (1989). Modeling the perception of tonal structure with neural nets. *Computer Music Journal*, 13 (3), 44-53.

Cooper, G. & Meyer, L.B. (1960). *The rhythmic structure of music*. Chicago: University of Chicago Press.

Fodor, J. & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28, 3-71.

Johnson-Laird, P.N. (in press). Jazz improvisation: a theory at the computational level. In P. Howell, R. West, and I. Cross (Eds.), *Representing musical structure*. NY: Academic Press.

Jordan, M. I. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. Proceedings of the Eight Annual Conference of the Cognitive Science Society. Hillsdale, N.J.: Erlbaum Press.

Knopoff, L. & Hutchinson, W. (1978). An index of melodic activity. *Interface*, 7, 205-229.

Lerdahl, E. & Jackendoff, R. (1983). *A generative theory of tonal music*, Cambridge: MIT Press.

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

Palmer, C. & Krumhansl, C.L. (1990) Mental representations of musical meter. *Journal of Experimental Psychology: Human Perception & Performance*, 16, 728-741.

Pollack, J. B. (1988). Recursive auto-associative memory: Devising compositional distributed representations. Proceedings of the Tenth Annual Conference of the Cognitive Science Society, Montreal, 33-39.

Pollack, J. B.(1990). Recursive distributed representations. *Artificial Intelligence*, 46, 1, 77-105.

Pressing, J. (1988). Improvisation: methods and models. In J. Sloboda (Ed.),*Generative processes in music: the psychology of performance, improvisation, and composition*. NY: Oxford University Press.

Rosenthal, D. (1989). A model of the process of listening to simple rhythms. *Music Perception*, 6 (3), 315-328.

Scarborough, D. L., Miller, B. O., & Jones, J. A. (1989). PDP models for meter perception. Proceedings of the Eleventh Annual Conference of the Cognitive Science Society. Hillsdale, N.J.: Erlbaum Press.

Schenker, H. (1979). *Free composition* (E. Oster, Trans.). NY: Longman

Steedman, M. (1982). A generative grammar for jazz chord sequences. *Music Perception*, 2, 52-77.

Todd, P. M. (1989). A connectionist approach to algorithmic Composition. *Computer Music Journal*, 13 (3), 29-43.