REDUCED MEMORY REPRESENTATIONS FOR MUSIC

Edward W. Large

Caroline Palmer

Jordan B. Pollack

The Ohio State University

Abstract

We address the problem of musical variation (identification of different musical sequences as variations) and its implications for mental representations of music. According to reductionist theories, listeners judge the structural importance of musical events while forming mental representations. These judgments may result from the production of reduced memory representations that retain only the musical gist. In a study of improvised music performance, pianists produced variations on melodies. Analyses of the musical events retained across variations provided support for the reductionist account of structural importance. A neural network trained to produce reduced memory representations for the same melodies represented structurally important events more efficiently than others. Agreement among the musicians' improvisations, the network model, and music-theoretic predictions suggest that perceived constancy across musical variation is a natural result of a reductionist mechanism for producing memory representations.

To appear: *Cognitive Science*

REDUCED MEMORY REPRESENTATIONS FOR MUSIC

1 Introduction

A common observation about musical experience is that some musical sequences are heard as variations of others. The tendency of listeners to hear musical variation has been exploited by composers and performers of various cultures and styles for centuries. Examples from Western music include “theme and variations” forms of classical and romantic music, and the improvisational forms of modern jazz. This problem is not specific to music; the problem of perceptual constancy in the face of physical variation is central to cognitive science. This problem has interested many researchers in the field of music cognition because the invariance of musical identity that characterizes the listener's experience is perceived across a wide range of differences in the surface content of the music (Dowling & Harwood, 1986; Lerdahl & Jackendoff, 1983; Schenker, 1979; Serafine, Glassman, & Overbeeke, 1989; Sloboda, 1985). To explain phenomena such as musical variation, most theorists rely on structural descriptions of musical sequences. Depending on the musical dimension(s) under consideration the nature of the description will vary, but each relies on some abstract system of knowledge representing the underlying regularities of a particular musical style or culture. Through experience with a musical style, listeners are thought to internalize characteristic patterns of rhythm, melody, harmony, and so forth, which are used to integrate and organize musical sequences, and afford experiences such as musical variation.

The problem that musical variation presents to theorists is best illustrated by an example. Consider the melodies of Figure 1. The melodies labelled A are the children's tunes “Hush little baby” (top), and “Mary had a little lamb” (bottom). The melodies labelled B are improvisations on these tunes, performed by pianists in an experiment described in this paper. Most listeners readily identify the B melodies as “variations” of the A melodies: listeners believe that the B melodies share an identity with the original melodies. However, one's listening to these examples or inspecting the musical notation will reveal that at the surface level these two sequences differ along a number of dimensions, including pitch content, melodic contour, and rhythm. Where is the similarity between these sequences? One possibility is that as listeners produce internal representations for musical sequences, they implicitly evaluate the structural importance of events. Thus, certain events may be more important than others in determining the relationships that listeners hear between the melodies and variations of Figure 1. The evaluation of relative importance allows listeners to create reduced descriptions of musical sequences that retain the gist of the sequences and at the same time reduce demands on memory. Other theorists have made similar proposals, which we refer to as reductionist theories of music comprehension (Deutsch & Feroe, 1981; Lerdahl & Jackendoff, 1983; Schenker, 1979). Reductionist theories propose to explain musical variation by positing a similarity of the underlying structures in related melodies.

Figure 1: A) Two original melodies, “Hush little baby” (top), and “Mary had a little lamb” (bottom). B) Improvisations on each tune, from Experiment 1.

This explanation for musical variation, that listeners produce reduced memory descriptions for musical sequences, poses three major challenges for theories of music cognition. The first is the problem of knowledge specification. Reductionist theories posit that an experienced listener assigns to a musical sequence a relative importance structure that is based on previously acquired information: information that is not necessarily present in the individual musical sequence. Therefore, the notion of reduction requires that listeners implicitly use style- and culture-specific musical knowledge in creating reduced descriptions. This type of structural description is extracted from the input with the aid of general knowledge about the roles that events play in a particular musical idiom, and is thought to reflect the statistical regularities of the particular musical culture or style in question (cf. Knopoff & Hutchinson, 1978; Palmer & Krumhansl, 1990). A great deal of effort has gone into the explicit identification of the musical knowledge required for listeners to produce reduced descriptions of musical sequences (Lerdahl & Jackendoff, 1983; Schenker, 1979). However, no mechanism has been proposed that specifies what form such knowledge might take or how this knowledge is put to use.

A second major problem is that of empirical validation. Do listeners hear musical sequences in terms of a hierarchical structure of relative importance? Some empirical support for this claim has begun to emerge in the literature. Earlier studies have examined the role of relative importance in perceptual phenomena such as judgements of musical completion and stability (Palmer & Krumhansl, 1987a; 1987b), and judgements of similarity (Serafine, Glassman, & Overbeeke,1989). However, little research has investigated the role of structural importance in musical performance tasks such as improvisation, in which performers are required to create musical variations.

Finally, a third major challenge remains for reductionist accounts of music comprehension: the specification of learning mechanisms. The fact that reduced descriptions require culture- and style-specific musical knowledge implies that a complete theory of musical reduction will also have a significant learning component. Krumhansl (1992) argues that listeners abstract and internalize underlying regularities through experience with musical patterns. These cognitive representations give rise to expectations and affect the stability of memory (Krumhansl, 1979; Krumhansl, Bharucha, & Castellano, 1982). Jones (1981) argues that listeners abstract and store “ideal prototypes” of musical styles, that lead to musical expectations. Unexpected events in music create interest, but are more difficult to recall (Jones, Boltz, & Kidd, 1982). However, little work has addressed the mechanisms by which listeners may learn the musical regularities or prototypes that are necessary for the identification of the underlying structure of specific musical sequences.

In this article, we discuss the problem of musical variation and its implications for the mental representation of musical sequences. In particular, we address each of these three challenges posed for reductionist theories. We propose a mechanism that is capable of producing reduced memory representations for music, based on sequences that are first parsed into constituent data structures. We then model the reduced descriptions using recursive distributed representations (Pollack, 1988, 1990), a connectionist formalism that allows us to represent symbolic data structures as patterns of activation in connectionist networks. We also describe an empirical study in which we collect and analyze musicians' improvised variations on three melodies. We compare the improvisations with predictions of structural importance based on reductionist accounts. The evidence from improvisational music performance addresses the validity of reductionist claims and their relationship to the problem of musical variation. It also provides us with empirical data to compare with the performance of the connectionist mechanism for producing reduced memory representations. Finally, we describe a computational experiment in which a neural network is trained to produce recursive distributed representations for the same three melodies used in the improvisation study. The network model demonstrates a form of learning, providing an example of how listeners may acquire intuitive knowledge through passive exposure to music that allows them to construct reduced memory representations for musical sequences. We test the network's ability to generalize: to produce reduced descriptions for a musical variation of a known melody, and for a completely novel musical sequence. An examination of the reduced descriptions reveals that the representations differentially weight musical events in terms of their relative importance, thus emphasizing some aspects of the musical content over others. Finally, we compare these results with the empirical study to address whether the network's differential weightings agree with the relative importance of events inferred from the improvisational music performances.

1.1 Theories of Memory for Music

As Dowling and Harwood (1986) have observed, the role of memory in listening to a piece of music is not unlike the role of memory in listening to a conversation. In order to understand what is being said at any given moment, one need not have perfect recall of the conversation up to that point; what is important is the overall meaning or gist of the previous conversation. Listening to a piece of music is similar in at least one important way. For even moderately complex pieces, most listeners do not literally remember every detail; instead, they understand a complex piece by a process of abstraction and organization, remembering its musical “gist”. Psychologists studying patterned sequence learning in the 1950's and 1960's made similar observations regarding individuals' abilities to perceive and memorize patterns in time. Two main theoretical proposals were advanced to explain the psychological findings: recoding theories, and rule-formation theories. The concept of information recoding, first introduced by Miller (1956), suggested that subjects presented with to-be-remembered sequences can reduce the amount of information to be retained by recoding, or chunking, subsets of more than one item into a single memory code. Researchers such as Estes (1972), Vitz and Todd (1969), and Garner and Gotwald (1968) argued that subjects assign codes to the subgroups of a sequence in order to reduce demands on memory, and these codes can be recalled and decoded on a later occasion to reconstruct the entire sequence. The principles proposed for grouping elements to produce codes were often perceptual; for example, Vitz and Todd (1969) suggested that runs of perceptually similar elements are cast into memory codes. However, the recoding view has been criticized for its reliance on perceptual regularities and for its inability to explain subjects' abilities to predict upcoming events in patterned sequences.

In contrast to recoding theories, rule-formation theories emphasized individuals' use of ordered vocabularies, or alphabets, and rules that apply to alphabetic properties. Several researchers (Jones, 1974; Restle, 1970; Simon & Kotovsky, 1963) proposed that subjects abstract serial relations and, using rule-based transformations such as repeat, transpose, complement, and reflection, generate cognitive data structures. The use of such transformations was thought to account for subjects' abilities to represent and predict unfolding serial patterns. Simon and Sumner (1968) proposed that listening to music could similarly be modelled as a process of pattern induction and sequence extrapolation, using alphabets and rule-based transformations such as *same* (repeat) and *next* (next element in the alphabet).

Both recoding and rule-formation theories, however, fail to explain the extraction of invariant identification in musical variation. To handle this and other challenges posed by musical experience, reductionist theories of music cognition posit cognitive representations that identify the structural importance of musical events (Deutsch & Feroe, 1981; Lerdahl & Jackendoff, 1983; Schenker, 1979). One of the most comprehensive of the reductionist theories is Lerdahl and Jackendoff's “Generative Theory of Tonal Music” (1983). The theory takes as its goal the description of the musical intuitions of listeners experienced with Western tonal music. This is accomplished using a combination of music-theoretic analyses, of which *metrical structure* and *time-span reduction* are the most relevant for our purposes.

The primary function of metrical structure analysis is to describe the sense of alternating strong and weak pulses that characterizes musical experience, called *metrical accent*. A metrical structure consists of beats: psychological pulses marking equally spaced points in time[1]. Stronger and weaker pulses form nested levels of beats. The larger the metrical level of the beat marking a temporal location, the stronger that beat location, as shown by the dots in Figure 2 for the melody “Hush little baby”. A second function of metrical structure is to mark the onsets of temporal chunks that, in combination with grouping rules, divide a musical sequence into rhythmic units called *time-spans*. The resulting *time-span segmentation* divides a piece into nested time-spans. It captures aspects of the piece's rhythmic structure, providing a constituent structure description for the entire musical piece, as shown by the brackets in Figure 2.

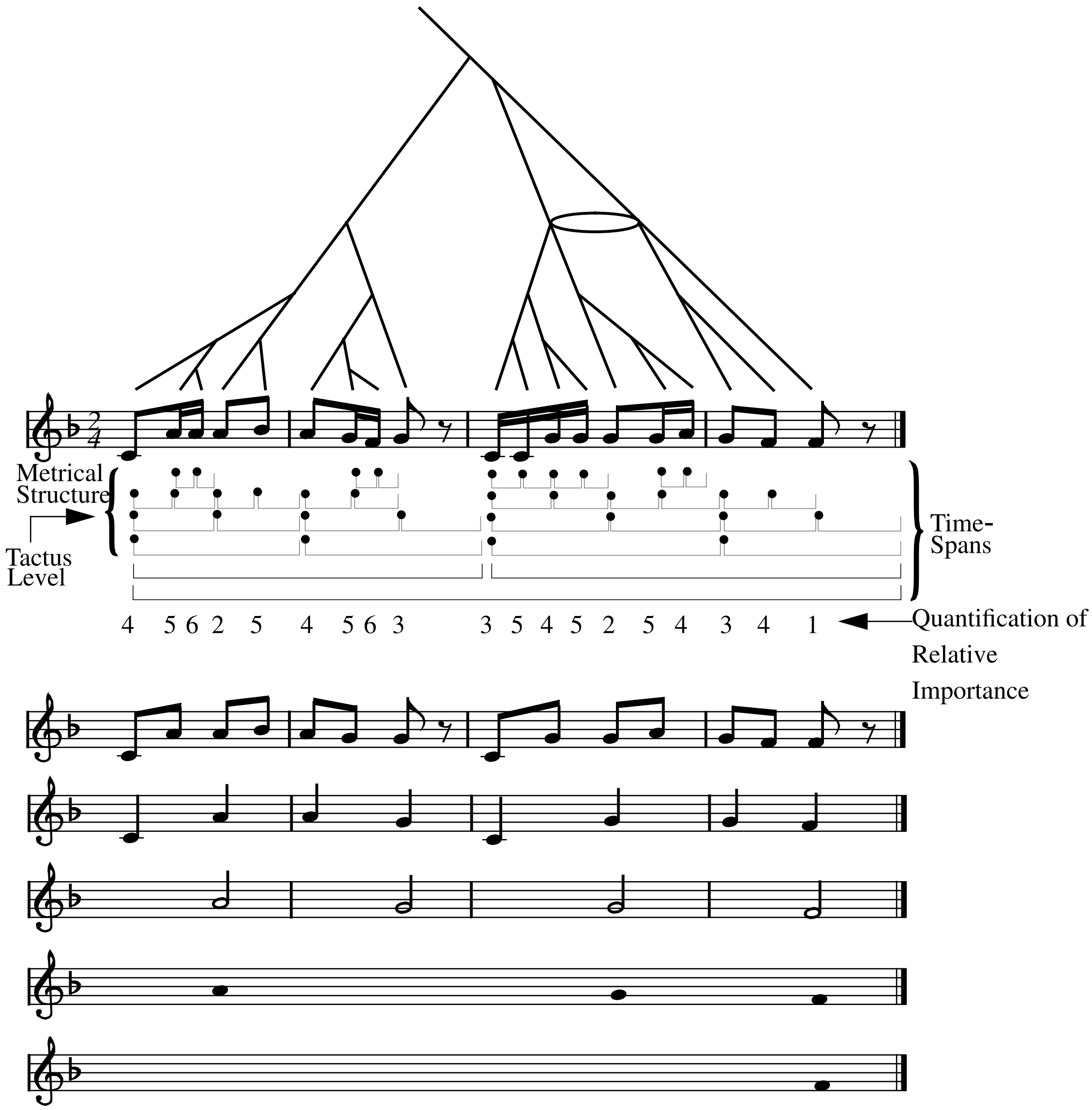


Figure 2. Analysis of "Hush little baby" following Lerdahl & Jackendoff (1983), showing metrical structure analysis, time-span segmentation, and time-span reduction. The original melody is shown in musical notation and the tree above it is the time-span reduction. The lower staves show the dominant events at each level of the time-span segmentation (marked in brackets). The metrical structure analysis is marked as rows of dots, and the quantifications of relative importance for each event are shown below the segmentation.

A time-span segmentation forms the input to a *time-span reduction* analysis, which organizes musical events into a structure that reflects a strict hierarchy of relative importance. Within each time-span a single most important event, called the head of the time-span, is identified. All other events in the time-span are heard as subordinate to this event. The time-span reduction assigns relative importance to each event according to rules that consider metrical accent in addition to melodic, harmonic, and structural factors. Thus, time-span reduction provides a unification of musical factors and predictions regarding which events listeners will perceive to be most important. Figure 2 shows a time-span reduction; the top musical staff shows the melody and the staves below show the heads for successively larger and larger time-spans. At each level, the less important event(s) of each time-span are eliminated, and a "skeleton" of the melody emerges. The tree above the top musical staff combines the information conveyed by the skeletal melodies with the information conveyed by the time-span segmentation. Its branching structure emphasizes structural relationships between levels of the reduction: that events of lesser importance are heard as elaborations of the more important events. The tree also identifies the structural ending of the musical passage, the cadence, indicated by the ellipse "tying" together two branches of the tree, as shown in Figure 2.

Reductionist theories can be applied to explain the perception of musical variation. Figure 3 compares the theoretical reduction of the original melody "Hush little baby" with a reduction of the improvised variation on this melody (from Figure 1). At the third skeletal level, the two reductions are identical. Lerdahl and Jackendoff's (1983) theory thus can be applied to predict an intermediate level of mental representation at which structural similarities are captured. These theoretical reductions can be quantified, as shown in Figure 2; the numbers correspond to the relative importance of each event described by the time-span reduction analysis. Each number is a count of the number of branch points passed in traversing the tree from the root to the branch that projects in a straight line to the event, inclusive. For instance, to calculate importance for the first note of the melody, count 1 for the root, 1 for a left turn, 1 for a branch point passed, and 1 for a second left turn. This final branch projects in a straight line to the event, and so counting stops[2]. According to this strategy, the smaller the number, the more important is the corresponding event. Metrical accents also make predictions of relative importance based on event location. These predictions can be quantified by counting the levels of beats that correspond to metrical predictions (the dots in Figure 2). The two measures are usually correlated because time-span reduction is partially based on metrical accent, but the time-span reduction adds information beyond metrical structure. We will compare both quantifications of relative importance and metrical accents, computed as in Palmer and Krumhansl (1987a, b), with the measures from improvisational music performance and with predictions derived from connectionist formalisms, described next.

1.2 A Reductionist Mechanism

In this section we address the issue of the mechanism by which time-span reductions may be computed, given a time-span segmentation as input. One difficulty with designing a mechanism based on Lerdahl and Jackendoff's (1983) theory lies in the specification of a relative weighting scheme for the set of rules that create reductions. A scheme has not yet been proposed that will work for every musical context. For complex musical pieces, one must enlist the aid of musical "common sense" in providing the proper weighting of musical considerations. A second problem regards learning. Reductionist theories assume that a great deal of musical knowledge is acquired as a result of experience with the musical culture or style in question. Empirical evidence suggests that a restructuring of mental representations for novel musical sequences may occur with as few as five or six exposures to a sequence (Serafine, Glassman, & Overbeeke, 1989). However, reductionist theories have not yet addressed the issue of how the musical knowledge necessary for the production of reduced descriptions is acquired.

A.

B.

Figure 3. The first three levels of reduction for “Hush little baby” and its improvised variation (from Figure 1). The reductions are identical at the third level.

One approach that offers a solution to these problems is the application of connectionist models, which learn internal representations in response to the statistical regularities of a training environment using general-purpose learning algorithms such as back-propagation (Rumelhart, Hinton, & Williams, 1986). The solution for musical variation offered by reductionist theories requires the representation of constituent structure, however, and connectionist models have been notoriously weak at representing constituent relationships such as those in language and music (Fodor & Pylyshyn, 1988). In fact, the lack of useful compositional representations has been an important stumbling block in the application of neural networks and other pattern-recognition techniques to the problems of cognitive science in general. One solution to this problem involves learning distributed representations for compositional data structures using a recursive encoder network. This connectionist architecture, known as Recursive Auto-Associative Memory (RAAM), has been used to model the encoding of hierarchical structures found in linguistic syntax and logical expressions (Chalmers, 1990; Chrisman, 1991; Pollack, 1988; 1990).

To produce a memory representation for a musical sequence with the RAAM architecture, we first parse the sequence to recover a compositional data structure that captures the sequence's time-span segmentation. Thus, the network's input represents a musical sequence and its constituent structure; it does not capture the relative importance information conveyed by metrical accent or time-span reduction. We then train a RAAM network to produce a distributed representation for each time-span described by this structure. For example, the sequence of musical events in Figure 4, “a b c”, may represented as the nested structure ((a b) c). A compressor network is trained to combine a and b into a vector R1, and then to combine the vector R1 with c into a vector R2. A reconstructor network is trained to decode the vectors produced by the compressor into facsimiles (indicated by the symbol ') of the original sets of patterns. In the example, the reconstructor decodes R2 into R1' and c', and R1' into a' and b'. Thus, the vector R2 is a *representation* for ((a b) c) because we can apply a reconstruction algorithm to R2 to retrieve a facsimile of the original sequence. It is a *distributed* representation because it is realized as a pattern of activation. It is a *recursive* distributed representation because its construction requires the network to recursively process representations that it has produced. The representations are *reduced descriptions* of musical sequences because the vector representation for an entire pattern is equal in size to the vector representation of a single event.

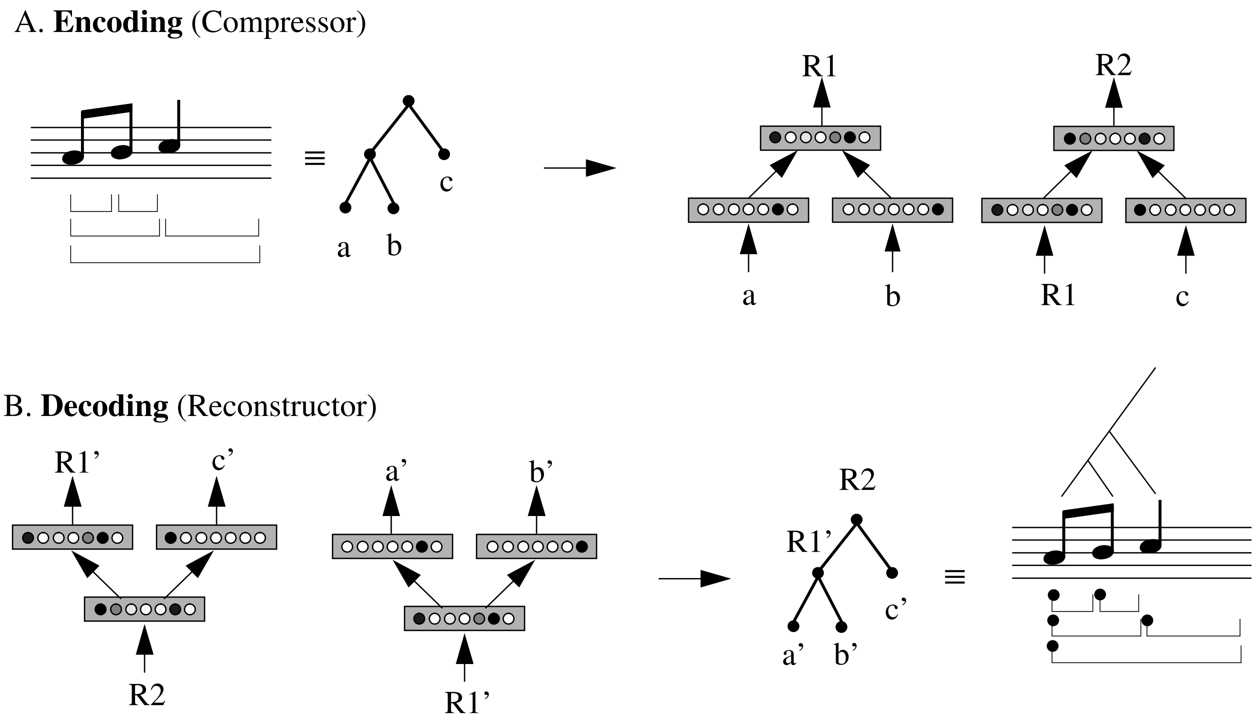


Figure 4. Encoding and decoding of a musical sequence by a RAAM network. A) Based on a vector representation for each event and a constituent structure analysis, the compressor combines the group (a b) into a single vector, R1, and then combines (R1 c) into a single vector R2. B). The reconstructor decodes the vector R2 to produce the facsimile (R1' c'). It then decodes R1' into (a' b').

The structures that the RAAM reconstructs are facsimiles of the original structures because the construction of a recursive distributed representation is a data compression process, which necessarily loses information. The network may reconstruct some events with lowered activation, and may fail to reconstruct other events entirely. The question we address regards which events will be reconstructed faithfully, and which will be lost or altered in the compression/reconstruction process. If, in the compression/reconstruction process, the network consistently loses information pertaining to less important events and retains information about more important events (i.e. as predicted by the music-theoretic analyses), then the network has also captured information that extends beyond pitch and time-span segmentation. The test is whether or not the network training procedure discovers the relative importance of events corresponding to metrical accent and time-span reduction.

If the network passes this test, then the use of recursive distributed representations to represent musical sequences may provide answers to some of the questions we have posed for reductionism. Reductions of musical sequences may be computed by a memory coding mechanism whose purpose is to produce descriptions for musical sequences that reduce demands on memory while retaining the gist of the sequences. This implies that the culture- and style-specific musical knowledge necessary for computing reductions is realized as a set of parameters (in a RAAM network, a set of weights) in the coding mechanism. The acquisition of this set of parameters can be viewed as the acquisition of the musical knowledge relevant to computing reductions.

This view of reduced memory representations for musical sequences has a number of advantages over other possible mechanisms. The vector representations produced by a RAAM for melodic segments are reduced descriptions of the sequence, similar to the “chunks” proposed by recoding theorists. However, the compressed representation for a sequence is more than just a label or pointer to the contents of a structure (cf. Estes, 1973); it actually *is* the description of its contents. Therefore, the numeric vectors produced by the network potentially contain as much information as the cognitive structures proposed by rule-formation theorists. Because the reduced descriptions are represented as neural vectors, they are suitable for use with association, categorization, pattern-recognition and other neural-style processing mechanisms (Chrisman, 1991). Such processing mechanisms could, for example, be trained to perform sequence extrapolation tasks (Simon & Sumner, 1968).

In the next section we address issues of empirical validation. Do humans weight melodic events in terms of relative importance? Can reductionist accounts explain the phenomenon of invariant identification across musical variation? We describe an empirical study of variations on melodies improvised by skilled pianists. We extract a measure of relative importance for each melody based on the improvisations. We compare these measures to reductionist predictions based on Lerdahl and Jackendoff's (1983) theory. In the following section, we describe a study in which we train a RAAM network to produce reduced descriptions for a set of melodies. We then test the network on the same three melodies on which pianists improvised in the empirical study. We measure the network's ability to produce representations for these melodies, including the ability to recognize melodic variation. We describe a method for determining the network's assignment of relative importance to individual events in the melodies, and compare the network findings with the empirical data and with the theoretical predictions. In the final section, we discuss the implications of reductionist theories for models of human learning and memory.

2 Empirical Investigation

Empirical evidence supporting the reductionist point of view has emerged in the literature. However, these early studies have dealt primarily with perceptual phenomena (Palmer & Krumhansl, 1987a; 1987b; Serafine, Glassman, & Overbeeke, 1989). The reductionist hypothesis also leads to predictions concerning music performance. For example, in musical traditions that employ improvisation, performers may identify the gist of a theme in terms of its structurally important events and use techniques of variation to create coherent improvisations on that theme (Johnson-Laird, 1991; Lerdahl & Jackendoff, 1983; Pressing, 1988). Therefore, it should be possible to identify the events of greater and lesser importance in a melody by collecting improvisations on that melody and measuring the events that are retained across improvisations. We use this rationale to identify structurally important events by asking performers to improvise variations on a melody, and we examine the variations for events altered or retained from the original melody.

A number of methods have been employed to elicit the structure of listeners' mental representations for musical sequences. For example, Palmer and Krumhansl (1987a; 1987b) asked subjects to listen to excerpts from a musical passage and rate how “good or complete” a phrase each excerpt formed. The rating was taken as a measure of the relative importance for the final event in each musical excerpt. Listeners' judgements of phrase completion at various points in a musical passage correlated well with predictions of each events' relative importance derived from Lerdahl and Jackendoff's (1983) time-span reductions (Palmer & Krumhansl, 1987a; 1987b). The nature of the musical task, however, was somewhat unnatural, because music is usually not presented in fragments. Additionally, the application of this paradigm to longer musical works is problematic. Serafine, Glassman, and Overbeeke (1989) asked listeners to judge the similarity between related melodies. Although this paradigm does not provide measures of importance for individual musical events, it does allow the assessment of reductionist claims within an ecologically valid task. The similarity judgements among melodies corresponded to the degree of relatedness predicted by a reductionist theory (Schenker, 1979), even when radical surface differences existed (such as in the musical harmony). This agreement increased with repeated hearings, indicating a significant role of learning in determining the structure of listeners' mental representations (Serafine, Glassman, and Overbeeke, 1989). Schenker's reductionist theory, although similar in spirit to Lerdahl and Jackendoff's proposal, is less specific in its description of rules and their applications, often requiring a trained analyst (Serafine, Glassman, and Overbeeke, 1989) to make judgements regarding theoretical predictions.

The experiment reported here is based on a paradigm described earlier (Large, Palmer, & Pollack, 1991). In this paradigm, musicians are presented with notated melodies and are asked to improvise (create and perform) simple variations on them. Improvisation in Western tonal music commonly requires a performer to identify some framework of melodic and harmonic events, and apply procedures to create elaborations and variants on them (Johnson-Laird, 1991; Steedman, 1988; also, see Pressing, 1988 for a review of improvisational models). Thus, improvisation of variations allows the musician freedom to determine which if any musical events should be retained from the original melody. This paradigm addresses the reductionist account by measuring musicians' intuitions about a particular melody within the context of a familiar task. This paradigm has an additional advantage in that it allows for the collection of individual ratings of importance for each event. Musical events that are viewed as structurally important should tend to be retained in improvised variations. Events viewed as less important (i.e., events that function as elaborations of important events) should be more likely to be replaced with different elaborations.

We measure the relative importance of each pitch event in the original melody by counting the number of times it was retained in the same relative temporal location across improvisations. Although this is a coarse measure of improvisation, it allows us to generalize across many aspects specific to music performance (including dynamics, phrasing, rubato, pedaling, etc.) and improvisation (including motific development, stylistic elaboration, etc.), and concentrate instead on those factors that reflect reductionist considerations.

The primary objective of this study was to extend earlier findings (Large, Palmer, & Pollack, 1991) that suggested that a musician's improvisations on a tune indicated an underlying reduced representation of the melody. According to our application of the time-span reduction hypothesis to improvisation, more important events (those retained across multiple levels of the time-span reduction) should be more likely than unimportant events to be retained in variations on a melody. Therefore, the number of individual pitch events retained in the musicians' improvisations should correspond to the theoretical predictions of structural reductions.

2.1 Method

2.1.1 Subjects

Six skilled pianists from the Columbus, Ohio community participated in the experiment. The pianists had a mean of 17 years (range of 12 to 30 years) of private instruction, and a mean of 24 years (range of 15 to 32 years) of playing experience. All of the pianists were comfortable with sight-reading and improvising. All were familiar with the pieces used in this study.

2.1.2 Materials

Three children's melodies (“Mary had a little lamb”, “Baa baa black sheep”, and “Hush little baby”) were chosen as improvisational material that would be familiar (well-learned) for most listeners of Western tonal music, to insure a well-established notion of relative importance for each event and to avoid learning effects. Additionally, these pieces were fairly unambiguous with regard to their time-span reductions.

2.1.3 Apparatus

Pianists performed on a computer-monitored Yamaha Disklavier acoustic upright piano. Optical sensors and solenoids in the piano allowed precise recording and playback without affecting the touch or sound of the acoustic instrument. The pitch, timing, and hammer velocity values (correlated with intensity) for each note event were recorded and analyzed on a computer.

2.1.4 Procedure

The following procedure was repeated for each piece. Pianists performed and recorded the melody, as presented in musical notation, five times. These initial recordings allowed each pianist to become acquainted with the improvisational material. With the musical notation remaining in place, the pianists were then asked to play five “simple” improvisations. The pianists were also asked to play five “more complex” improvisations, which are not discussed here. All performances were of a single-line melody only; pianists were instructed not to play harmonic accompaniment. All pianists indicated familiarity with all of the musical pieces.

2.2 Results

2.2.1 Coding Improvisations

Each improvisation was coded in terms of the number of events retained from the original melody, to develop a measure of relative importance for each event. The following procedure applied to the coding of each improvisation. First, the improvisation was transcribed by two musically trained listeners, who agreed on the transcriptions. Next, sections of the improvisation were matched to sections of the original. For most improvisations this was straightforward; for two of the improvisations, sections that repeated in the original melody ("Baa baa black sheep") were rendered only once in the improvisation, and these were doubled for purposes of analysis. Finally, individual events of the improvisation were placed into correspondence with the original. If only the pitch contents and rhythm changed (meter and mode remained the same), as in most of the improvisations, this process was straightforward: events were placed into correspondence by metrical position. In the case of mode change (for example, the flatted third is substituted for the major third in a major to minor mode shift), substitutions were counted as altered events. In the case of a meter change, metrical structures were aligned according to the onsets of each measure and half-measure, and events were then placed into correspondence by temporal location. Those events whose pitch class was retained in the correspondence between original melody and variation were coded as "hits" and received a score of 1; those events whose pitch class was altered (or for whom no event corresponded in the improvisation) were coded as "misses" and received a score of 0. For example, if a quarter note "C" were replaced with four sixteenth notes "C-B-C-B" beginning at the same metrical location, the "C" would be coded as a hit. If, however, the "C" had been replaced with "B-C-B-C", the "C" would be coded as a miss. Thus, only deletions and substitutions of events in the original melody affected the number of hits.

The number of hits for each pitch event in the original melody was summed across the five improvisations for each performer. Figure 5 shows the mean number of retained events across performers for each melody. To rule out the possibility that events in the original melody were altered at random, or that performers simply added events to create improvisations, an analysis of variance (ANOVA) on performers' mean number of retained events by event location was conducted for each melody. Each of the three ANOVAs indicated a significant effect of event location (Melody 1: $F(25,125) = 4.02$, $p < .01$; Melody 2: $F(52, 260) = 6.64$, $p < .01$; Melody 3: $F(18, 90) = 7.76$, $p < .01$). Thus, performers were more likely to retain some melodic events than others across improvisations. The factors influencing the number of retained events at each location were further investigated in the following analyses.

2.2.2 Comparison with Theoretical Predictions

Both metrical accent and time-span reduction make predictions about relative importance based on event location. Correlations between improvisation measures and both sets of theoretical predictions for each melody are summarized in Table 1. First, the correlation between the number of pitch events retained and the quantified metrical accent predictions for each event location was significant for each melody ($p < .05$). Improvisation measures were next compared with predictions from the time-span reduction analysis for each melody, obtained by quantifying the number of branch points passed in the tree, from root to terminal branch, as shown in Figure 2. Correlations

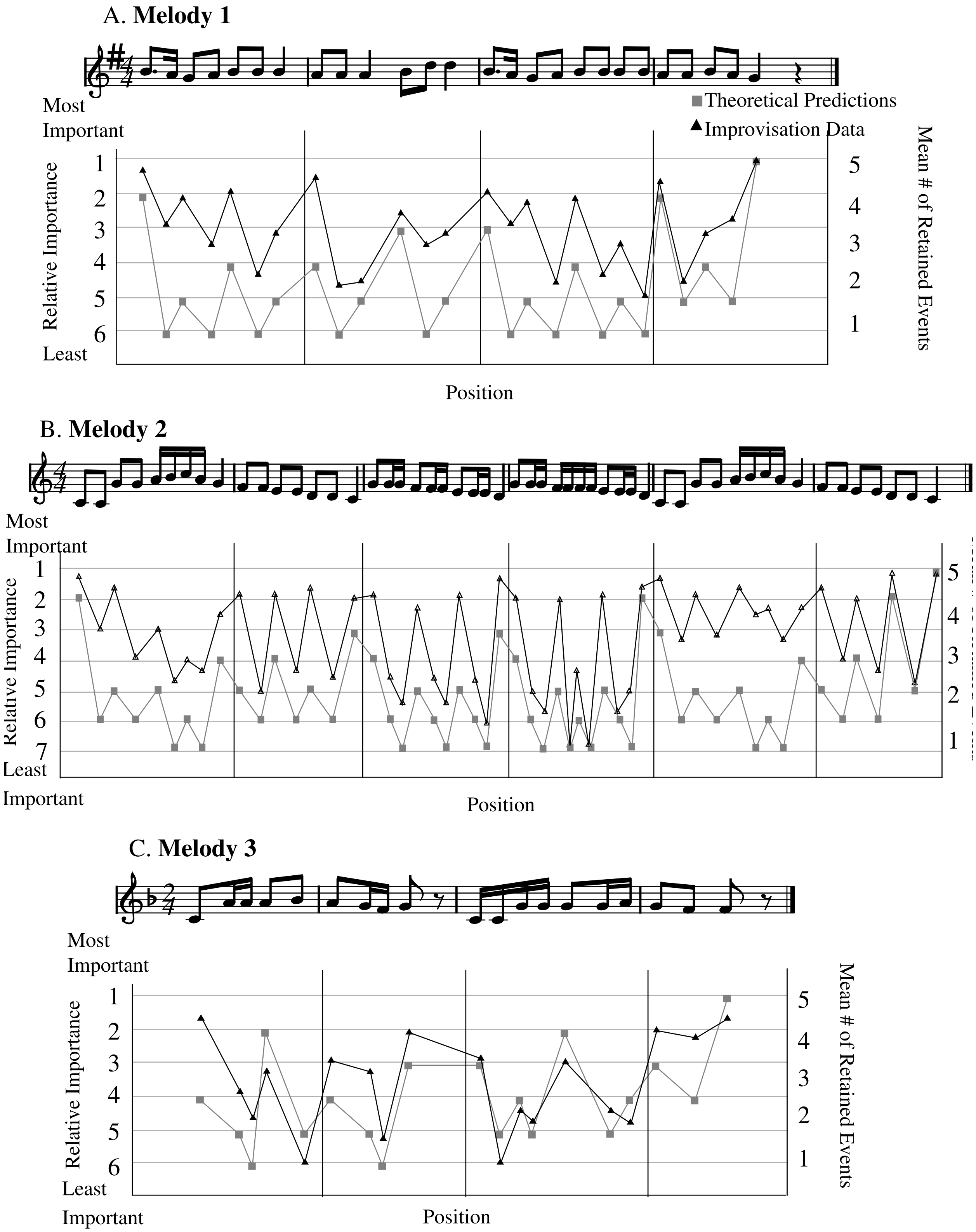


Figure 5. Theoretical and empirical measures of relative importance for melodies A ("Mary had a little lamb"), B ("Baa baa black sheep"), and C ("Hush little baby").

between the number of pitch events retained and the quantified time-span reduction predictions were also significant for each melody ($p < .05$). Figure 5 shows the time-span quantifications along with the improvisational data, indicating that those events predicted to be most important according to the time-span reduction tended to be retained across improvisations.

Table 1: Squared correlation coefficients for theoretical predictions and improvisation-based measures.

	Melody 1: (Mary)	Melody 2: (Baa)	Melody 3: (Hush)
Metrical Accent Predictions	.63*	.80*	.78*
Time-Span Predictions	.76*	.79*	.67*
Semi-Partial (metrical accent removed)	.42*	.30*	.21

* - p <.05

To insure the predictive power of the time-span reduction beyond metrical accent (on which time-span reductions are partially based), the improvisation measures were correlated with time-span reduction predictions after the effects of metrical accent were partialled out. These semi-partial correlations, also shown in Table 1, were significant ($p<.05$) for melodies 1 and 2, indicating that time-span reduction did contribute information beyond metrical accent. The semi-partial correlation was not significant for melody 3 ($p=.37$), indicating that in this case correlation of improvisation measures with the time-span reduction analysis was largely due to the effects of metrical accent.

2.3 Discussion

Musicians' improvisations of variations on simple melodies provided strong support for the reductionist hypothesis. Performers tended to retain certain events in each melody, and used improvisational techniques to create variations around those retained events. In addition, the music performances agreed with reductionist predictions of which events were relatively important in these simple melodies. Furthermore, the findings for two of the three melodies indicate that musical factors specific to time-span reductions played an important role in musicians' improvisation of variations.

The relatively high contribution of metrical structure to the improvisations based on the third melody (“Hush little baby”) may indicate a qualitative difference between the performers' intuitions and the theoretical predictions for this piece. For example, the improvisations often retained the first event of measure 1 (as seen in Figure 5), an indicator of its relative importance, disagreeing with the theoretical weighting of this event. This may be due to the salience of the large initial pitch interval, or it may be a general primacy effect (making the first few events more likely to be retained regardless of reductionist considerations). The performances also disagreed with the predictions at the structural ending; all events in measure 4 were retained relatively often. Alternatively, this could be accounted for as a recency effect.

These discrepancies emphasize the difficulty of providing a relative weighting for a set of rules that determine the reductionist structure of mental representations. For example, the particular order in which a subset of rules is applied can lead to different weightings of constituents. However, the improvised performances do show general agreement with the theoretical predictions of time-span reduction. This is the first demonstration, to our knowledge, that the musical factors incorporated in the reductionist theory (Lerdahl & Jackendoff, 1983) can account for the structure of performers' mental representations for musical improvisations. We next compare these measures with the reduced memory descriptions generated by a connectionist network for the same three melodies.

3 Network Model

In this section we describe an implementation of the Recursive Auto-Associative Memory (RAAM) architecture, for producing reduced memory descriptions of musical sequences. A RAAM network is trained on a corpus of simple melodies and is then tested in two ways. First, we examine the network's ability to accurately compress and reconstruct a test set of three tunes. Second, we examine the structure of the representations produced by the network. The RAAM network is an example of a mechanism that applies knowledge about a musical genre to the task of producing reduced memory descriptions for specific musical sequences. This knowledge is captured as a set of weights that the network extracts directly from a training environment (the corpus of melodies), thereby accounting for learning.

The network experiment has two goals. The first is to measure the performance of the RAAM network using a *well-formedness* test (Pollack, 1990). For a given input melody, the compressor network creates a reduced description. The reconstructor network is then applied to the reduced description to retrieve its constituents. If the reconstructed sequence matches the input melody, either exactly or within some tolerance, then the reduced description is considered to be well-formed. The well-formedness test can also be used to measure the ability of a RAAM network to generalize, by testing the network's performance on novel sequences. In this experiment, we examine the performance of the network on a test set of three melodies: *known*, *variant*, and *novel*. The *known* melody is one of the melodies presented to the network during a training phase. Performance on this melody establishes a baseline of the network's ability to correctly encode melodies. The *variant* melody is a variation of material presented to the network in the training phase, and the *novel* melody is a melodic sequence not related in any obvious way to the material presented in the training phase. If the network is able to generalize from the examples presented in the training phase, then it should be able to produce well-formed reduced descriptions for one or both of the *variant* and *novel* melodies as well as for the *known* melody.

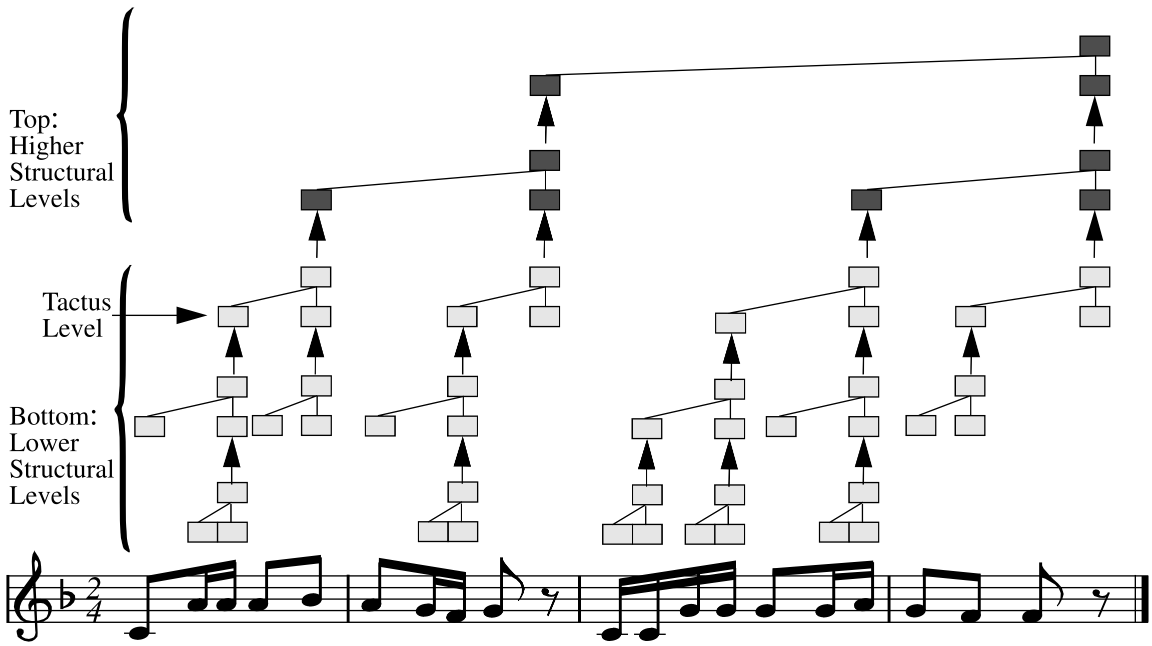
The second goal is to determine the structure of the representations produced by the network. The network is provided with a time-span segmentation for each melody. The question we ask is: Will the network be able to take advantage of this information about temporal structure to preserve musical regularities that are systematically related to this structure? Our prediction is that the network's reduced descriptions will differentially weight constituent events, and furthermore, that this differential weighting will agree with both the improvisation measures and the theoretical predictions (Lerdahl & Jackendoff, 1983) regarding the relative importance of events. This prediction is based on the observation that RAAM networks learn a data compression algorithm that is tailored to the statistical regularities of a training set. If our training set is adequately representative of the statistical characteristics of simple Western tonal melodies, the network should be able to make use of this information, and we should see significant levels of agreement among network, theoretical, and empirical measures.

3.1 Network Implementation

The RAAM architecture uses a connectionist substrate of fully-connected feed-forward neural networks to produce recursive distributed representations (Pollack, 1990). For example, to encode binary trees with k-bit patterns as the terminal nodes, the RAAM compressor would be a single-layer network with 2 k-unit input buffers and one k-unit output buffer. The RAAM reconstructor would then be a single-layer network with one k-unit input buffer and 2 k-unit output buffers. The input and output buffers are required to be the same size because the network is used recursively: the output of the network is fed back into the network as input. During training, the compressor and reconstructor are treated as one standard three-layer network ($2k$ inputs, k hidden units, and $2k$ outputs) and trained using an auto-associative form of back-propagation (Rumelhart, Hinton, & Williams, 1986; Cottrell et.al., 1988), in which the desired output values are simply the input values. To create the individual training patterns for the network, the structures that make up the training set are divided into two-element groups (for example: (a b) or (R1 R2)).

Two special issues arise in designing a RAAM network for encoding musical structure. First, we must determine a constituent structure for each musical sequence that will specify how events are presented to the network input buffers, as shown in Figure 6 (top). As discussed above, we will use the time-span segmentation, a nested constituent description that captures aspects of the sequence's rhythmic structure (Lerdahl & Jackendoff, 1983). In the simple melodies used in this study, the time-spans at smaller constituent levels (less than a measure) were “regular” (Lerdahl & Jackendoff, 1983); that is, they were aligned with the locations of strong metrical beats. Therefore, the lower levels of time-span segmentation were determined by the metrical structure. Grouping rules (Lerdahl & Jackendoff, 1983) contributed to determine time-span segments at constituent levels larger than the single measure. Each encoding produced by the network is the representation of a time-span (and its events) at some level in the time-span segmentation. Once an encoding has been produced, temporal information is implicitly managed by the recursive structure of the decoding process, also shown in Figure 6 (bottom). As decoding proceeds, the output codes represent smaller and smaller time-spans (at lower and lower levels), until, at the termination of the decoding process, a single pitch event is output and the temporal location of that event is uniquely determined. Thus the temporal structure of each melody plays an important part in network processing; the reduced descriptions of the melody in Figure 6 capture temporal structure in a way analogous to Lerdahl and Jackendoff's (1983) hierarchically nested time-spans, shown for the same melody in Figure 2.

A. Encoding (Compression)



B. Decoding (Reconstruction)

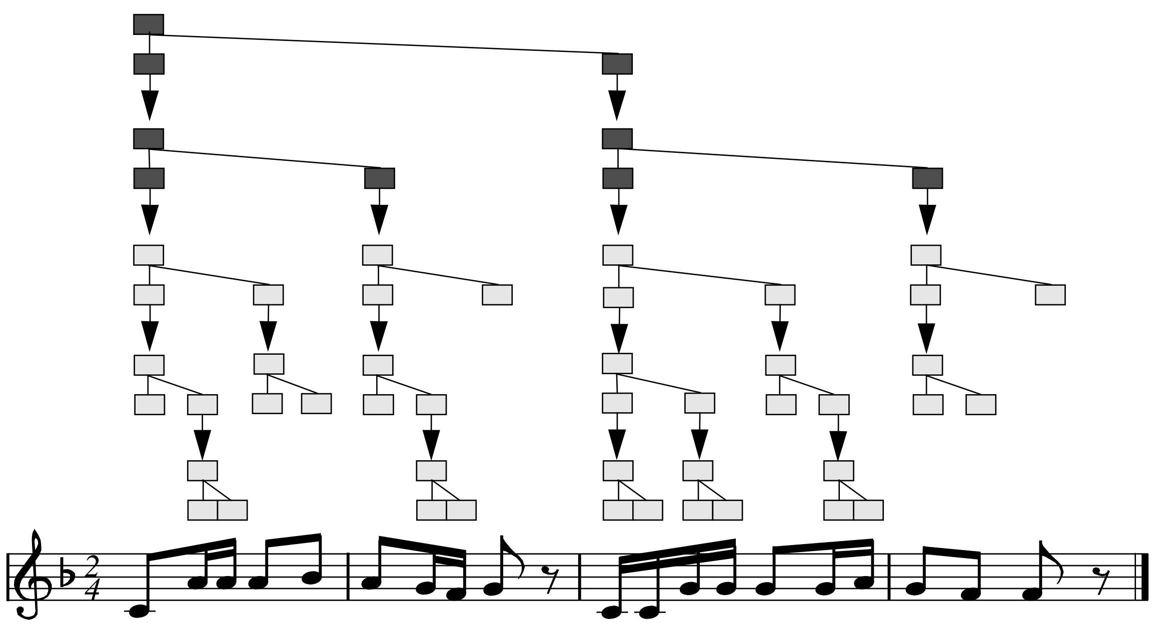


Figure 6. Simplified schematic of the modular RAAM used in the network experiment. A) In the encoding diagram, time flows from left to right and bottom to top. B) In the decoding diagram, time flows from top to bottom and left to right.

Second, we must specify the representation of pitch events to be encoded by the RAAM. The different pitches in each melody are represented as binary feature vectors (on or off). We chose a “local” representation of pitch class; 7 units represented the seven pitch classes of the diatonic scale in Western tonal music. We also added two units to represent melodic contour. One unit means 'up' from the previous event, the other means 'down', and turning both units off means no contour change. This representation nominally captures octave equivalence and pitch height but makes no further assumptions regarding the psychophysical components of pitch, as other connectionist researchers have done (cf. Mozer, 1991). More sophisticated encoding strategies may prove useful for certain musical applications (cf. Large, Palmer, & Pollack, 1991), but for the purposes of this study we sought to minimize inductive biases that would be introduced by more complex coding schemes.

Two modifications to the RAAM architecture are necessary to encode Western tonal melodies such as those in our training set. First, existing applications of the RAAM architecture have only accurately handled tree structures that are 4-5 levels deep. However, the 25 training melodies used in this study contain constituent structure hierarchies 6-7 levels deep, which expand to more than 1000 individual training patterns. Previous experiments found that this training set size outstrips the capacity of a RAAM network that contains a reasonably small number of hidden units (Large, Palmer, & Pollack, 1991). We adopt a method here of scaling up the basic architecture by having one RAAM network recursively encode lower levels of structure, and then passing the encodings it produces to a second RAAM that encodes higher levels of structure, as shown in Figure 6. This method, known as *modular* RAAM (Angeline, 1990; Sperdutti, 1992), allows us to build recursive encoders that can handle trees with many hierarchical levels by using multiple networks that each contain fewer hidden units. This form of training, however, violates one of the original design decisions of RAAM: that the terminals be recognizable as binary strings so that it is clear when to terminate the decoding process (Pollack, 1990). We address this problem by adding an extra unit to each RAAM module, that is trained as a terminal detector. This allows us to determine a) when to pass a code from the higher-level RAAM network module to the lower-level RAAM module during decoding, and b) when to interpret a code produced by the bottom module as a pitch event[3].

The second modification addresses ternary groups common in Western tonal music; binary branching structures are not sufficient to capture musical groupings that often consist of three elements. To handle both pairs and triples, we might create a network with three input buffers, only two of which would be used to encode binary segments. However, this would lead to the situation shown in Figure 7A, in which a triple with a rest in the middle is indistinguishable from a pair. Instead, a network with four input buffers can encode both duple and triple segments and distinguish among them, as shown in Figure 7B. Here buffer 1 corresponds to the first event of any group, buffer 3 corresponds to the second event of a binary group, and buffers 2 and 4 correspond to the second and third events, respectively, of a ternary group. In order to properly interpret the output of these buffers at decoding, we add four extra units at the output. The network is trained to turn on an output unit when the corresponding buffer's output is to be used; otherwise the contents of the buffer are ignored, and trained with a don't-care condition (Jordan, 1986).

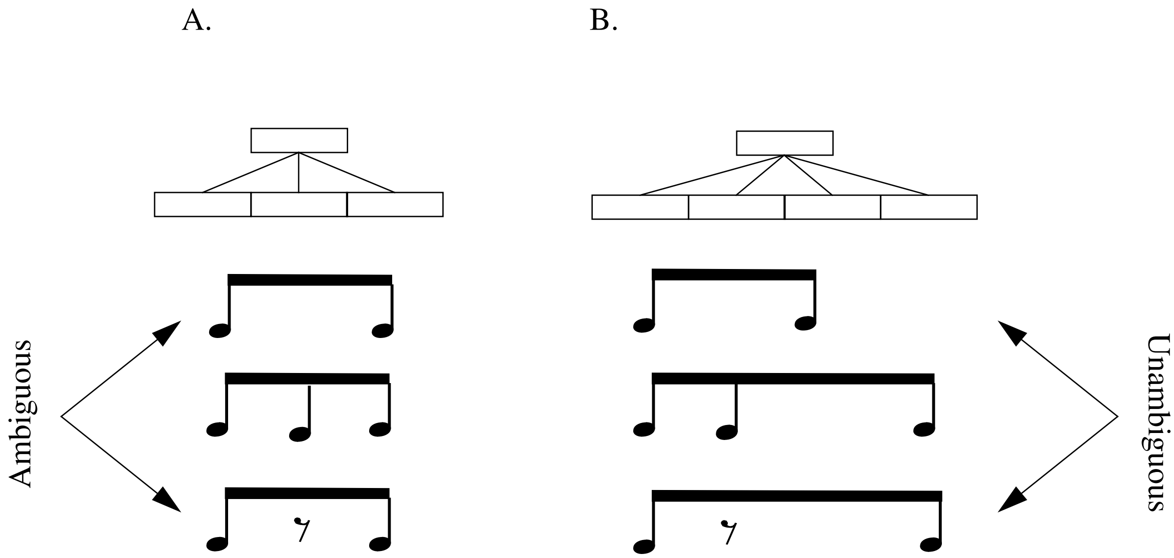


Figure 7. Proposed buffering for encoding duple and triple grouping structures. A) Three buffers cannot discriminate between a group of two events and a group of three events in which the middle event is a rest. B) Four input buffers can make the discrimination.

3.2 Method

3.2.1 Training the Network

Twenty-five simple children's melodies (listed in the Appendix) were chosen as a training set because they provided a simple, natural musical case for study. The tunes comprised eighteen unique melodies; five of these eighteen melodies had variations in the training set. Each melody was between 4 and 12 measures in length, with a time-signature of 2/4, 3/4, 4/4, 6/8, or 12/8. The tunes provided constituent structures six to seven levels deep, in which either binary or ternary groups appeared at each level. Although the pitch event representations required only 9 bits (7 pitch class units and 2 contour units), we used 35 units, allowing 26 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-cares (Jordan, 1986) on output. As described above, the two RAAM modules each required four input buffers, and each resulting module had 140 input units, 35 hidden units, and 148 output units.

The first module was trained on the bottom 2-3 levels of the trees, such that the input corresponded to metrical levels up to and including coding of the “tactus”, or beat level (see Figure 2). Therefore, the representations that emerged from the lower RAAM (the output) corresponded to time-spans with a length of one half-note for binary groups, and one dotted half-note for ternary groups (see Figure 6). The second module was trained on the upper 3-4 levels of the trees, corresponding to larger structural levels of the melodies. This division of labor allowed the modular architecture to approximately balance the learning load between the two modules, measured by the number of unique training patterns. The two modules were trained simultaneously, with the bottom module's output providing the input for the top module. Rather than the network being exposed to the entire training set of twenty-five melodies, four melodies were chosen randomly from the training set and forward-propagated in each training cycle; then error was back-propagated through the network (cf. Cottrell & Tsung, 1991). This method allowed a faster running time for the large training set. Because the length of the tunes in the training set (and therefore the number of individual training patterns) varied for each cycle of backpropagation, the learning rate was set to 0.7 divided by the number of training patterns seen on that cycle. Momentum was set to 0.5 and weight decay to 0.0001. Training lasted for 1300 cycles, at which point the error associated with the test set of melodies reached a minimum value.

3.2.2 Testing the Network

We tested the network's performance in two ways. First, well-formedness tests assessed the ability of the network to accurately compress and reconstruct each melody, and thereby revealed the basic representational capacity of the network. Second, tests of representational structure assessed the relative weighting of constituents on an event-by-event basis, and thereby revealed the nature of the representational strategy developed by the network.

The network was tested on a set of three melodies: a *known* melody, a *variant* melody, and a *novel* melody, shown in Figure 8. These were the same three melodies used in the empirical study of improvisation; the names used here denote the particular relationship of each melody to the network training set. The *known* melody, “Mary had a little lamb”, occurred in the training set (Appendix 16A; all melodies are shown in the key of C). The network's performance on this melody is representative of the network's performance on familiar (learned) melodies. The *variant* melody, “Baa baa black sheep”, did not occur in the training set; however, four closely related variations of this melody (18A-D) did occur in the training set. The local structure (duration patterns and melodic contour of individual measures) of the variant melody was very similar to that of training set melodies 18 C and D. The global structure (three 2-measure phrases with similar melodic and harmonic implications) of the variant melody was similar to training set melodies 18A and B. Performance on this melody indicates the ability of the network to account for simple melodic variation, because the network was required to recombine familiar local structures (individual measures) in novel global contexts (different melodies) that shared structural features with known melodies. The *novel* melody, “Hush little baby”, did not occur in the network training set, nor was it closely related to any of the melodies in the training set. Network performance on this melody indicates the ability of the network to perform a type of generalization different from that required for melodic variation: the ability to represent novel musical sequences at local levels of structure, as well as the ability to combine novel local structures in novel global contexts.

3.3 Results

3.3.1 Tests of Well-Formedness

Each melody was reconstructed by the decoder network from the recursive distributed representation produced by the compressor. Errors in the reconstructed melody took the form of additions (the network reconstructed an event that was not present in the original melody), deletions (the network failed to reconstruct an event that was present in the original melody), and substitutions (the network reconstructed an event incorrectly in the same position). These errors correspond to two aspects of the network's performance. The first is whether or not the network correctly reconstructs the time-span segmentation, which determines the rhythm (duration pattern) of the output sequence. The second is whether or not the network correctly reconstructs the contents of the output vectors, which correspond to the pitch contents of the output sequence. We combined these considerations into a single measure of network performance by calculating whether the network gives the correct output for each possible temporal location. Each of the melodic sequences in the test set had a smallest durational value of a sixteenth-note, and no reconstruction produced any smaller temporal values. Therefore, we based our error measure on the number of sixteenth-note locations in each piece. There were 64 (16x4) sixteenth-note locations in the known melody, 96 (16x6) in the variant melody, and 32 (8x4) in the novel melody. Given our coding scheme, the chance estimate for percentage of events correct at each location is 1/16, or 6.25%, based on 16 possible outcomes: seven pitch classes times two contour changes (up or down) plus a repeated pitch and a rest.

As an approximate measure of the network's ability to correctly compress and reconstruct constituent structures, we calculated average performance on the training set melodies. We measured performance at two points in the time-span segmentation for each melody. First, the network's ability to compress and reconstruct time-span segments with only three levels of recursive nesting – corresponding to a time-span of one half note for binary groups – was examined. Network performance in reconstructing training set melodies with three levels was 92%. Next, the network's ability to compress and reconstruct time-span segments that corresponded to entire melodies, with 6-7 levels of recursive nesting, was examined. Here the network's performance was 71%. Thus, the representations captured lower-level structures quite faithfully, whereas at global levels of structure, the representations began to lose sequence details. (The network's performance at the lower levels will always be more accurate than at the global levels because more data compression at global levels results in greater susceptibility to loss of information.)

In order to better understand the network's performance, we examined the reconstructions for the three test melodies in detail, which are shown in Figure 8. The reconstruction of the known melody, “Mary had a little lamb”, gives an indication of network performance on melodies learned in the training set. As described above, we first examined the reduced descriptions produced by the lower-level RAAM module for the known melody (subsequences of events up to the level of half notes; lowest 3 levels of hierarchical nesting). In this reconstruction, the network made a single error, adding an event in the third measure, for a performance of 98%. Reconstruction at the whole tune level (all 7 levels of hierarchical nesting) resulted in four errors, giving an overall performance of 94% (60/64) for this melody, which was significantly better than chance (binomial test, $p < .01$). This reconstruction was better than the average for training set melodies, probably due to the fact that two instances of this melody (16A and B) occurred in the training set. Note also that the network's reconstruction of the first measure of this melody at the whole tune level resembles the duration pattern of melody 16B; however, the reconstructions of 16A and 16B did differ (the network was able to differentiate between them).

The reconstruction of the *variant* melody, “Baa baa black sheep”, gives an indication of the network's performance on simple variations of learned melodies. The reconstruction produced by the lower-level RAAM module for subsequences corresponding to half-notes (3 lowest levels of hierarchical nesting) resulted in performance of 92% (88/96). The network successfully learned the lower-level details because most of these surface features were present in the training set (see melodies 18C and D). The network's reconstruction at the whole tune level (all 7 levels of nesting) resulted in fifteen errors, for a performance of 84% (81/96), again significantly better than chance ($p < .01$). The reconstruction of this melody at (only) the whole-tune level was the same as its reconstruction of "Twinkle twinkle little star", one of the four related melodies in the training set (18A), for which its performance was 98% (94/96 events). This is a not a surprising result, but it is revealing to note that the network reconstructed melody 18A rather than one of the other related training set melodies (18B-D). As Figure 8 shows, the half-note level representations preserved local structure. The ability to exploit constituent structure, combined with the use of a recursive encoding strategy, allowed the network to rely upon structural similarities at the whole-tune level, rather than melodic and rhythmic features at lower levels, in determining the representation of this melody.

The reconstruction of the *novel* melody, "Hush little baby", is indicative of the RAAM's ability to encode novel sequences. Again, we first examined the reduced descriptions produced by the lower-level RAAM module for subsequences of the original melody by encoding groups of events only up to the level of half-notes (3 levels of hierarchical nesting). Figure 8 shows the reconstruction from the reduced descriptions for each half-note of the tune. The lower-level reconstructions produced ten errors, for a success rate of 69% (22/32), again significantly better than chance ($p < .01$). Seven of the ten errors occurred in the third measure, and the other three measures of the tune were reconstructed rather faithfully. At the whole tune level there were seventeen errors, for a performance of 47% (15/32), which is significantly better than chance ($p < .01$), but overall, the reconstruction is rather poor (there are only 19 events in the original tune). It is interesting to note that the rhythm was reconstructed well (27/32, or 84%), but very few pitch events were reconstructed correctly (3/19, or 16%). Thus the network's representation of this melody at the whole tune level was not well-formed, and generalization to this novel sequence was better at the lower levels of the hierarchy.

3.3.2 Tests of Representational Structure

Next, we analyzed the structure of the distributed representations to determine the relative contributions of individual events. One method is to directly examine the representation vectors to determine the function of individual hidden units. Little information can be retrieved from recursive distributed representations of this size, however, because of their complexity (Pollack, 1990). As an alternative approach, the "certainty" with which the network reconstructed each event of the original sequence was measured by computing the distance between the desired (d) and obtained (o) vector representations at each sequence location (i). This analysis considered only those output units that represent pitch class (ignoring contour), consistent with the analysis of the improvisations. To compensate for the fact that some events were added and others deleted in the reconstructions, we considered only the locations in the reconstructions for which pitch vectors should have been output. Thus, only deletions and substitutions of events from the original melody affected this measure, as in the empirical study above. A similarity measure was defined,

$$sim(d, o) = 1 - \left(\sqrt{\sum_{i=1}^{n} (d_i - o_i)^2}\right) / n,$$

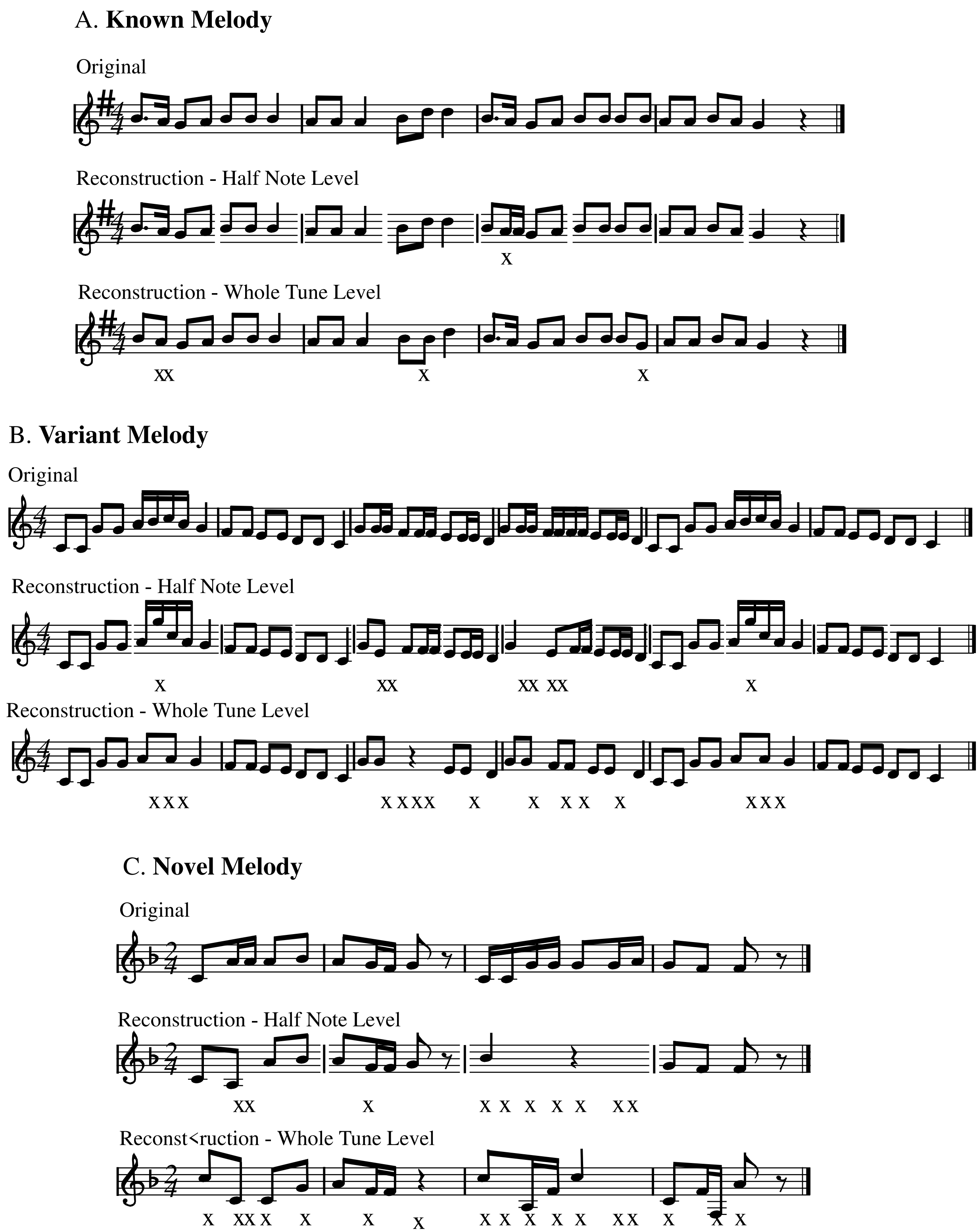


Figure 8. Original melodies and network reconstructions: A) "Mary had a little lamb" (Known), B) "Baa baa black sheep" (Variant), C) "Hush little baby" (Novel). Each melody was reconstructed from several codes (the half-note level RAAM), and from a single code (the whole-tune level RAAM). X's denotes failures in network reconstructions.

that ranged from 0 (most different) to 1 (identical), and represented the probability that desired pitch events occupied the appropriate positions in the original sequence, based on the network representation. Thus, sequence locations at which this measure was smallest were locations at which the network was most likely to make a reconstruction error. These probabilities were then interpreted as predictions of relative importance for each event in the distributed representation.

The probability measures of relative importance at the whole-tune level were correlated with the musical improvisation data, as summarized in Table 2. The correlations were large for the known ($p < .10$) and variant ($p < .05$) melodies, but not for the novel melody. This was not surprising because the novel melody did not have a well-formed distributed representation at the whole-tune level. However, when the novel melody was reconstructed from the reduced descriptions corresponding to the half-note level of the tune (shown on the bottom of Figure 8), the resulting correlation approached significance ($p < .10$).

Table 2: Squared correlation coefficients for network reconstructions.

	Known (Mary) Whole Tune	Variant (Baa) Whole Tune	Novel (Hush) Whole Tune	Novel (Hush) Half-Note
Improvisation Data (# events retained)	.35*	.64**	.10	.40*
Metrical Accent Predictions	.39**	.55**	.24	.45**
Time-Span Predictions	.39**	.64**	.25	.52**
Semi-Partial (metrical accent removed)	.14	.35**	.27	.29

*- p <.10

** - p <.05

Next, we compared the network measures of relative importance with the quantifications of theoretical predictions, as shown in Table 2. The correlations with time-span reduction predictions were significant for the known and variant melodies and for the measure-level reconstruction of the novel melody ($p < .05$) but not for the whole-tune-level reconstruction of the novel melody. The correlations with metrical accent predictions also were significant for each melody ($p < .05$). We therefore correlated the network measure with time-span reduction predictions after metrical accent was partialled out. The semi-partial correlation was not significant for the known or novel melodies, but was significant for the variant melody ($p < .05$), indicating some ability of the network to extract structure beyond metrical accent.

3.4 Discussion

We have demonstrated a mechanism, Recursive Auto-Associative Memory, that is capable of producing recursive distributed representations for musical sequences. A general learning algorithm, backpropagation, extracted sufficient information from a training set of 25 simple melodies to produce reduced descriptions of known, variant, and novel sequences. The reconstructions of melodies produced by the network were fairly accurate, but did not retain all of the details. The network produced reduced memory representations that preserved the important structural features of the sequences.

We first investigated the performance of the network using the RAAM well-formedness test. In all three test cases, the network failed to reconstruct some events, reconstructed other events incorrectly, and occasionally added some that were not present in the original sequence. However, three sources of evidence suggested that the representations successfully captured the major structural features of the melodies. First, the reconstructions were faithful to the rhythm of the original melodies, even in the case of the novel melody. Second, the network correctly reconstructed most of the pitches in the original melodies. Third, the events on which the network made reconstruction errors tended to be the less important events, as shown by the correspondence of network predictions of relative importance with theoretical predictions and improvisational data.

The network performed best on familiar (learned) melodies, and differentiated between subtle variations of the same melody (16A and 16B). We also tested the ability of the network to generalize: to represent both a variant of a learned melody and a truly novel melody (one unrelated to the learned melodies). The performance of the network in reconstructing the variant melody showed how the network handles simple melodic variation. This melody shared local structure with training set melodies 18C and D, and the network's lower-level codes (up to the half-note level) preserved this structure. At a global level, the compression/reconstruction process followed the attractor (a known path) for another melody with which the variant shared global structure (18A). The network also identified the important pitch events in the variant, indicated by the fact that network measures of relative importance for this melody correlated strongly with the time-span reduction predictions. Comparison with the empirical data from improvisations supported the conclusion that the network successfully identified events interpreted as major structural features by musicians. Overall, these results indicate the ability of the network to exhibit a limited but important form of generalization.

The findings for the novel melody indicated that the network still performed well at lower levels of structure in handling unlearned sequences; it produced well-formed memory representations for the three lowest levels of the constituent structure. At higher levels of structure, however, the network failed to generalize, reproducing the correct rhythm but incorrect pitches for this melody. This aspect of performance may be due to the learning environment, which may not have provided a rich enough set of patterns at higher levels of structure.

The information retained by the network in the compression/reconstruction process agreed well with music-theoretic predictions of the relative importance of musical events. The limited size of the training and test sets make it difficult to say precisely why the agreement occurred; however, the time-span segmentation used as input to the network was related to the music-theoretic predictions. The network used this information about rhythmic structure, coded as position in a fixed input buffer, to learn representations that retained musically important events and major structural features. For instance, the network may have learned metrical accent by weighting the first element of lower-level time-spans (which aligned with strong metrical beats) more heavily than others. The relative importance predictions, however, were based on more complex rhythmic relationships. To learn relative importance, the network may have learned other stylistic factors. For example, the RAAM network may have learned that the last event in each sequence was predictable - it is always the tonic. Thus, the network appears to have extracted some relationships beyond metrical accent, and did so strictly on the basis of the regularities in the training set. The network was forced to distill musical regularities such as these from the training set in response to two opposing pressures: 1) to retain as much information about each sequence as possible, and 2) to compress the information about each sequence into a pattern of activation over a small number of units.

Finally, and most importantly, we demonstrated the psychological plausibility of this approach to creating reduced memory representations for music. Certain events dominated the structure of the reduced descriptions by virtue of the fact that they had the greatest probability of being correctly reconstructed by the network. The events that dominated the network's reduced descriptions were precisely those events most important in the mental representations for these melodies measured by the musical improvisations and posited in the theoretical reductionist predictions. These findings indicate that the RAAM coding mechanism produced reduced descriptions for musical sequences that implicitly weighted events in each sequence in terms of their relative structural importance. This is a major finding because it supports the psychological plausibility of recursive distributed representations as an approach to modelling human memory. Combined with the network's overall performance in reconstruction, these findings suggest that the reduced memory representations successfully captured the structure of musical sequences in ways similar to the mental representations underlying improvisational music performance.

4 General Discussion

We began with the problem of musical variation: how do listeners and performers judge certain musical sequences to be variations of others? We have argued that musicians make such judgements based not on characteristics of the surface structure of sequences, but instead based on similarity of reduced memory representations formed from the sequences. We have provided empirical support for the reductionist account of musical memory from a study of improvisational music performance, in which pianists tended to retain structurally important events across improvisations on simple melodies. The improvisations corresponded well with predictions from a reductionist account of music perception. The link between performance and perception may be the nature of memory, that highlights the musical gist at the expense of other musical features. These reduced memory structures may be based on knowledge extracted from passive exposure to many musical patterns from the same style or culture. Evidence to support this claim was provided by a connectionist network model that learned to produce reduced memory descriptions for simple melodies. In this section, we explore the relationship between the musical improvisation findings and our reductionist approach to modeling musical memory. We discuss the problem of the recovery of rhythmic structure in music, and its relationship to the general problem of the recovery of constituent structure in sequence processing. We then compare our network model with other connectionist models that have been proposed for music processing.

4.1 Musical Improvisation and Reduced Memory Structures

The structure of memory greatly influences the content of musical improvisations. Improvisation 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, 1991; Steedman, 1982). An important aspect of the application of the reductionist view to improvisation is that it relieves a potentially heavy burden on short-term memory. Instead of remembering each element in a musical sequence, only a reduced set of elements must be retained, from which improvisations can be generated. Johnson-Laird (1991) argues that, given an appropriate memory representation of musical structure, acceptable jazz improvisations can be generated by relatively simple computational processes, a constraint imposed by the demands of real-time processing. Thus, the relationships we have identified between the improvisational music performances, computational model, and reductionist theories of music cognition may result from similar representational requirements.

The view of musical memory that we propose is also closely related to Pressing's (1988) account of musical improvisation. In his view, the performer's mental representation for a musical sequence consists of event clusters: arrays of objects, features, and processes with associated cognitive strengths. Improvisation consists of the generation of novel sets of event clusters based on a set of improvisational goals applied to the schematic representation of a theme, given a memory of the event clusters previously generated (Pressing, 1988). Our connectionist network produced reduced memory descriptions for “clusters” of events in which structurally important events dominated. Furthermore, the variations improvised by the performers in our experiment were related in terms of similar underlying representations of the musical theme, and events with greater cognitive strength occurred more often in the improvisations. The improvisation measures of relative importance correlated strongly with the theoretical predictions and network measures, suggesting that reductionist memory representations capture the cognitive strength of events in improvisations of musical variations.

4.2 Temporal Structure and Constituent Structure

The model of memory representation for music that we have proposed relies on constituent structure that is not computed by the network architecture, but is available as input to it. Other connectionist researchers in music have assumed that knowledge about constituent structure should be extracted from a training set using general learning algorithms and then explicitly or implicitly dealt with by a network memory coding mechanism (Mozer, 1992; Todd, 1991). Our assumption is based on the theory that the constituent structures most relevant to music cognition are closely related to rhythmic structure (Lerdahl & Jackendoff, 1983). In this section we briefly discuss the relationship between rhythmic structure and constituent structure in music, and its relationship to general issues of constituent structure in temporal sequence processing.

Typically, listeners are provided with many cues or markers to rhythmic structure in music. Musical signals contain complex forms of temporal organization, including periodic structure on multiple time scales (Cooper & Meyer, 1960; Jones & Boltz, 1989, Large & Kolen, in press; Lerdahl & Jackendoff, 1983; Palmer & Krumhansl, 1990; Yeston, 1976) and temporal variation in expressive performance (Drake & Palmer, 1993; Palmer, 1989; Shaffer, Clarke, & Todd, 1985; Sloboda, 1983; Todd, 1985). The temporal information present in individual musical sequences affords the perception of rhythmic organization, including metrical and grouping structure. For example, mechanisms for the perception of metrical structure have been proposed that use only event inter-onset times and event accent information as input (Essens & Povel, 1985; Large & Kolen, in press; Longuet-Higgins & Lee, 1982). Additionally, expressive timing variations in musical performance provide cues for the perception of metrical and grouping structure (Palmer, 1988; Sloboda, 1983; Todd, 1985). Thus, the perception of metrical and grouping structure may not have to rely heavily on factors such as learned knowledge of melodic regularities. The perception of these forms of rhythmic organization combine to form the basis for the time-span segmentation (Lerdahl and Jackendoff, 1983), the constituent structure that we have assumed as input to our model.

Our network made use of this input information to produce representations that captured two kinds of musical structure. First, the network represented the time-span segmentation of the musical sequences. It used the time-span segmentation to adapt its processing strategy at each level, compressing and reconstructing groups of either two or three elements, to serve as an efficient encoder of predetermined structure. This is an important result because connectionist models are notoriously deficient at representing constituent structures as rich as those found in music. Next, the reduced descriptions captured the theoretically predicted and empirically observed “relative importance of musical events”. To accomplish this, the network used time-span segmentations to learn stylistic regularities that are systematically related to rhythmic structure. Such regularities, by definition, can only be captured by understanding a melody in relationship to knowledge of other melodies in that style.

Our assumption that constituent structure would be available as input to a memory process may be too strong. Previously learned patterns, such as cadences, can also influence listeners' perception of constituent structure (Lerdahl & Jackendoff, 1983). It seems likely that the perceptual processes responsible for segmenting a musical sequence into constituents and the representational processes responsible for encoding and/or recognizing constituents must interact. Reduced memory representations, such as those produced by RAAM networks, could mediate this interaction by facilitating the recognition of familiar patterns. The RAAM formalism provides a criterion by which we can measure the stability of any constituent's representation (the well-formedness test). In some cases, perceptual processes may simply provide the representational mechanism with groupings of events; in other cases, representational processes may effectively “choose” groupings of events that yield stable or familiar representations. Thus, we envision mutually supporting roles for the perception of temporal organization and the formation of memory representations for melodies.

4.3 Comparison with Other Connectionist Models

Other connectionist researchers have explored issues of sequential structure in music with discrete-time recurrent network architectures (Bharucha & Todd, 1989; Mozer, 1991; Todd, 1991). Recurrent networks are often trained to predict the next event of a sequence, given a memory of past events in the same sequence (Mozer, 1993), rather than being explicitly trained to develop a representation for an entire sequence. Recurrent networks are appealing for a number of reasons. First, recurrent networks are simple; they process a sequence one event at a time and assume no complex control mechanism. Second, recurrent networks are capable in principle of capturing arbitrary sequential and/or temporal relationships, including metrical structure, grouping structure, and even time-span reduction. Third, recurrent neural networks provide natural models of musical expectations for future events (Bharucha & Todd, 1991). Finally, recurrent networks can be used for musical composition. By connecting a network's output units to its input units, novel sequences are generated that reveal what has been learned about musical structure (Todd, 1991; Mozer 1991).

Recurrent networks have demonstrated some limitations in accomplishing these tasks, however. Recurrent networks have difficulty capturing relationships that span long temporal intervals, as well as relationships that involve high-order statistics (Mozer, 1993); thus they have had difficulty capturing the global structure of musical sequences (Todd, 1991; Mozer, 1991). In addition, at least one study suggests that the ability of discrete-time recurrent neural networks to learn simple temporal relationships, such as would be required to extract metrical structure, is also limited (McGraw, Montante, & Chalmers, 1991). In attempts to make recurrent networks more sensitive to the global structure of music, augmented versions of recurrent architectures have been proposed. One proposal is to build recurrent networks with hidden units that have various different constants of temporal integration. This allows the network to retain memory of past events more efficiently, and has resulted in networks with improved sensitivity to global structure (Mozer, 1992). Another proposal is to train hierarchically cascaded recurrent networks to explicitly extract and represent constituent structure (Todd, 1991). Neither of these proposals, however, (explicitly) uses temporal structure information to develop an emergent sensitivity to the relative importance of musical events, as our model does.

In principle, recurrent networks can represent any temporal relationships (e.g., metrical structure). In practice, however, discrete-time recurrent neural networks trained with backpropagation have not learned the temporal relationships that are most relevant for music. One reason may be because recurrent networks are not typically given material that offers much information for the recovery of temporal structure. Another reason may be that the architectures and learning algorithms employed are not biased toward discovering the temporal structures, such as metrical structure, to which humans are sensitive. For instance, the perception of rhythmic structure in music may be predicted primarily from duration and accent information present in individual musical sequences (Large & Kolen, in press; Longuet-Higgins & Lee, 1982; Palmer & Krumhansl, 1990). If recurrent networks were biased toward making use of relevant information about temporal structure, as is our network architecture, then they may be more likely to capture temporally relevant relationships such as relative importance.

4.4 Limitations and Future Work

The approach that we have described here has some limitations that highlight the need for further work. One regards the choice of musical materials in this study. The use of musical materials as simple as nursery tunes leads to some difficulties in interpreting the network findings. For instance, it is not clear whether the network's representational capability at global structural levels was limited by the network architecture or by the choice of training materials. In addition, the use of material as multifaceted as music means the relationship between metrical accent and time-span reduction predictions of importance were not controlled; the restriction to a small set of musical materials makes it difficult to determine how the network learns relative importance measures independently of metrical structure or how one might model these structural relationships in more complex improvisational forms of music. Thus, it is difficult to say precisely what structural relationships our model is capable of learning. Further study might use training and test melodies that control for interactions among structural relationships (cf. Elman, 1991).

Another possibility for further work concerns the choice of neural network architecture. One of the constraints of the RAAM architecture is the requirement of an external stack control mechanism for handling intermediate results during encoding and decoding (Pollack, 1988, 1990). In addition, the model requires a fixed-structure input buffer to make use of relevant temporal information. Although this buffer design is sufficient for handling simple melodies, a more complex buffer design would probably be necessary for melodies in which metrical structure and grouping structure may be misaligned or "out of phase" (Lerdahl & Jackendoff, 1983). Recurrent network architectures might be altered to exploit temporal information without necessarily entailing the restrictive design constraints of the RAAM architecture.

5 Conclusions

The phenomenon of musical variation can be explained by positing mechanisms that compute memory reductions. The reduced memory descriptions computed here for musical melodies resulted from encoding and decoding mechanisms that compressed and reconstructed the original sequences. These mechanisms led to reduced descriptions similar to those predicted by reductionist theories of music. This type of memory representation abstracts and summarizes sections of musical material, extracting what Dowling and Harwood (1986) refer to as the "gist" of a musical sequence. The reduced representations are suitable for manipulation by other neural-style processing mechanisms, and therefore may be useful for modelling musical tasks such as sequence extrapolation, structure recognition, and musical improvisation. A general learning algorithm (backpropagation) provided an example of how the knowledge relevant to computing reduced memory descriptions may be extracted from a learning environment, addressing an important challenge to reductionist theories. These findings support reductionist theories of music comprehension, but suggest that the computation of musical reductions is not an end in itself; rather, it is a natural result of the construction of memory representations for musical sequences.

Most importantly, we have demonstrated the psychological plausibility of reductionist theories of music comprehension, by comparing evidence from improvisational music performance with a model of reduced memory representations and with theoretical predictions regarding the relative importance of musical events. The fact that musical events were weighted similarly in musicians' choices of events to retain in improvisations, network encodings of the same melodies, and theoretical predictions of relative importance suggests that recursive distributed representations capture relevant properties of humans' mental representations for musical melodies.

References

Angeline, P. J., & Pollack, J. B. (1990). Hierarchical RAAMs. Technical Report 91-PA-HRAAMS, Laboratory for Artificial Intelligence Research, The Ohio State University, Columbus, OH.

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

Chalmers, D. (1990). Syntactic transformations on distributed representations. *Connection Science*, *2*, 53-62.

Chrisman, L. (1991). Learning recursive distributed representations for holistic computation. *Connection Science*, *3*, 345-366.

Cottrell, G. W., Munro, P. W., & Zipser, D. (1987). Image compression by back propagation: A demonstration of extensional programming. In N.E. Sharkey (Ed.), *Advances in Cognitive Science* (Vol. 2) (pp.305-321). Chichester, England: Ellis Horwood.

Cottrell, G. W., & Tsung, F. S. (1991). Learning simple arithmetic procedures, In J. A. Barnden and J. B. Pollack (Eds.), *High Level Connectionist Models*. Norwood, NJ: Ablex Publishing.

Deutsch, D., & Feroe, F. (1981). Internal representation of pitch sequence in tonal music. *Psychological Review*, *88*, 503-522.

Dowling, W. J., & Harwood, D. L. (1986). *Music cognition*. San Diego: Academic Press.

Drake, C., & Palmer, C. (1993). Accent structures in music performance. *Music Perception*, *10*, 343-378.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*, 179-211.

Essens, P. J. & Povel, D. (1985) Metrical and nonmetrical representation of temporal patterns. *Perception and Psychophysics*, <u>37</u>, 1-7.

Estes, W.K. (1972). An associative basis for coding and organization in memory. In A.W. Melton & E. Martin (Eds.), *Coding processes in human memory* (pp.161-190). NY: Halsted.

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

Garner, W. R., & Gottwald, R. L., (1968). The perception and learning of temporal patterns. *Quarterly Journal of Experimental Psychology*, *20*, 97-109.

Johnson-Laird, P. N. (1991). Jazz improvisation: a theory at the computational level. In P. Howell, R. West, and I. Cross (Eds.), *Representing musical structure*. London: Academic Press.

Jones, M.R. (1974). Cognitive representations of serial patterns. In B.H. Kantowitz (Ed.), *Human information processing*. NY: John Wiley & Sons.

Jones, M. R. (1976). Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological Review*, *83*, 323-335.

Jones, M.R. (1981). Music as a stimulus for psychological motion: Part I: Some determinants of expectancies. *Psychomusicology*, *1*, 34-51.

Jones, M. R., Boltz, M., & Kidd, G. (1982). Controlled attending as a function of melodic and temporal context. *Journal of Experimental Psychology: Human Perception & Performance*, *7*, 211-218.

Jordan, M. (1986). Serial order. Technical Report 8604, Institute for Cognitive Science, University of California at San Diego, La Jolla, CA.

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

Krumhansl, C.L. (1979) The psychological representation of musical pitch in a tonal context. *Cognitive Psychology*, *11*, 346-384.

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

Krumhansl, C. L., Bharucha, J. J., & Castellano, M. A. (1982). Key distance effects on perceived harmonic structure in music. *Perception & Psychophysics*, *32*, 96-108.

Large, E. W. & Kolen, J. F. (in press) Resonance and the perception of musical meter. *Connection Science*.

Large, E. W., Palmer, C., & Pollack, J. B. (1991). A connectionist model of intermediate representations for musical structure. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, (pp. 412 - 417), Hillsdale, N.J.:Erlbaum Press.

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

Longuet-Higgins, H. C. & Lee, C. S. (1982) The perception of musical rhythms. *Perception*, 11, 115-128.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81-97.

Mozer, M. C. (1991). Connectionist music composition based on melodic, stylistic and psychophysical constraints. In P. M. Todd and D. G. Loy (Eds.), *Music and connectionism* (pp.195-211). Cambridge: MIT Press.

Mozer, M. C. (1992). Induction of multiscale temporal structure.In R. P. Lippmann, J. Moody, & D. S. Touretsky (Eds.), *Advances in Neural Information Processing Systems 4*, (pp. 275-282), San Mateo, CA: Morgan Kaufman.

Mozer, M. C. (1993). Neural net architectures for temporal sequence processing. In A. Weigand N. Gershenfeld (Eds.), *Predicting the future and understanding the past* (pp. 243-264). Reading, MA: Addison-Wesley.

Palmer, C. (1988). Timing in skilled music performance. Unpublished doctoral dissertation, Cornell University, Ithaca, NY.

Palmer, C. (1989). Mapping musical thought to musical performance. *Journal of Experimental Psychology: Human Perception & Performance*, *15*, 331-346.

Palmer, C., & Krumhansl, C. L. (1987a). Pitch and temporal contributions to musical phrase perception: Effects of harmony, performance timing, and familiarity. *Perception and Psychophysics*, *41*, 505-518.

Palmer, C., & Krumhansl, C. L. (1987b). Independent temporal and pitch structures in determination of musical phrases. *Journal of Experimental Psychology: Human Perception & Performance*, *13*, 116-126.

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. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, (pp. 33-39), Hillsdale, N.J.:Erlbaum Press.

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

Pressing, J. (1988). Improvisation: methods and models. In J. Sloboda (Ed.), *Generative processes in music: the psychology of performance, improvisation, and composition* (pp.129-178). NY: Oxford University Press.

Restle, F. (1970). Theory of serial pattern learning: Structural trees. *Psychological Review*, *77*, 481-495.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation, In D. E. Rumelhart & J. L. McClelland (Eds.). *Parallel distributed processing* (pp.318-362). Cambridge: MIT Press.

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

Schmidhuber, J. (1992). Learning unambiguous reduced sequence descriptions. In R. P. Lippmann, J. Moody, & D. S. Touretsky (Eds.), *Advances in Neural Information Processing Systems 4*, (pp. 291-298). San Mateo, CA: Morgan Kaufman.

Serafine, M. L., Glassman, N., & Overbeeke, C. (1989). The cognitive reality of hierarchic structure in music. *Music Perception*, *6*, 347-430.

Simon, H. A., & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review*, *79*, 369-382.

Simon, H. A., & Sumner, K. (1968). Pattern in music. In B. Kleinmuntz (Ed.), *Formal Representation of Human Thought* (pp. 219-250). New York: Wiley.

Sloboda, J.A. (1983). The communication of musical metre in piano performance. *Quarterly Journal of Experimental Psychology*, *35*, 377-396.

Sloboda, J. A. (1985). *The musical mind*. Oxford: Oxford University Press.

Sperdutti, D. (1993). Representing symbolic data structures using neural networks. Unpublished doctoral dissertation, University of Pisa, Pisa, Italy.

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

Todd, P. M. (1991). A connectionist approach to algorithmic composition. In P. M. Todd and D. G. Loy (Eds.), *Music and connectionism*. Cambridge: MIT Press.

Vitz, P.C. & Todd, T.C. (1969). A coded element model of the perceptual processing of sequential stimuli. *Psychological Review*, *76*, 433-449.

Acknowledgments

This research was partially supported by NIMH grant 1R29-MH45764 to Caroline Palmer, and by ONR Grant N00014-92-J115 to Jordan B. Pollack. This paper was completed while the second author was a Fellow at the Center for Advanced Study in the Behavioral Sciences. We are grateful for financial support from NSF SES-9022192. We thank Fred Lerdahl for his comments on our stimulus materials and analyses and John Kolen, Gregory Saunders, and David Stucki for comments on an earlier draft. Reprint requests should be addressed to Edward W. Large, Department of Computer and Information Science, 2015 Neil Avenue, or to Caroline Palmer, Psychology Department, 1885 Neil Avenue, The Ohio State University, Columbus, OH 43210.

Reference Notes

[1] Beats themselves are often marked by the onset of acoustic events, but the sensation of beat can also occur when no event is physically present. When you tap your foot, or snap your fingers along with a piece of music, you are physically marking one particular level of beats from the metrical hierarchy called the tactus. According to Lerdahl and Jackendoff (1983), beats of the tactus level are always present in the perception of music. Beats marking smaller temporal levels are present only when acoustic events are present to mark them.

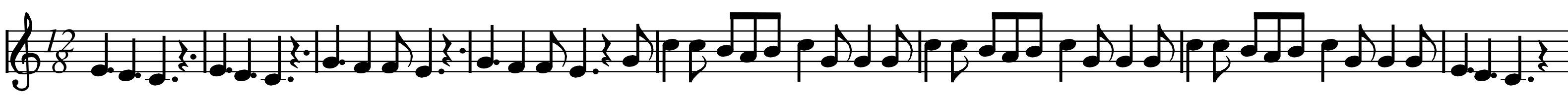
[2] The branch leading to the first event of a cadence is not counted as a branch point because it is considered structurally to be “part of” the final event (Lerdahl & Jackendoff, 1983). For example, to calculate importance for the first note of measure three, count 1 for the root, 1 for a branch point passed, 0 for a left turn (because this branch is tied), and 1 for a second left turn.

[3] Knowing when to terminate decoding is equivalent to determining the level of the time-span segmentation to which a melodic event corresponds. In general, levels of constituent structure in the network's input will correspond to levels of time-span segmentation and metrical accent; however, the network must learn the level of constituent structure at which to end the decoding process.

Appendix: Network Training Set

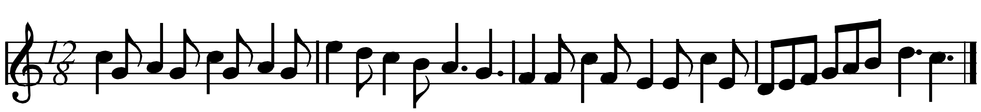
1 

2 

3 

4 

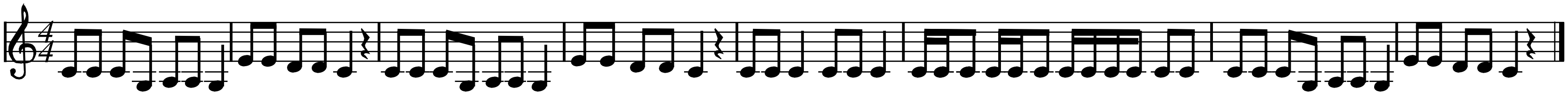
5 

6 

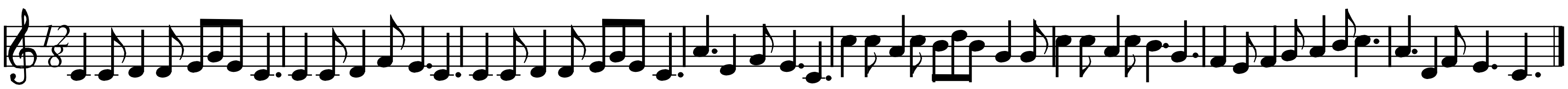
7A 

B 

8



9



10

11

12

13

14

15A

B

16A

B

17A

B

18A

B

C

D