Musical Pleasure through the Unification of Melodic Sequences

Bruce F. Katz
School of Cognitive and Computing Sciences,
University of Sussex
Brighton BN1 9QH UK
brucek@cogs.susx.ac.uk

Abstract

The purpose of this paper is to extend an earlier theory of pleasure associated with harmonic sequences to melodic sequences. The theory stated that the sequences that will be pleasurable will be the ones that allow a coherent transition from one mental state to the next. This can be measured in a connectionist model by noting the strength of the activation boost in a competitive layer categorizing the elements of the sequence. It is suggested that a similar mechanism will work for melodic sequences if two conditions are met. First, the melodic sequences must be represented such that two sequences judged to be similar by the ear have an overlapping distributed representation. Next, a mechanism must be posited to separate melodic sequences into significant groups. The results of a network accomplishing these tasks are presented.

Introduction

This paper will attempt to answer a question which is fundamental to musical cognition, and, by extension, to cognition in general. Why is music pleasurable? By what principles can one derive its power to stir, and to move, but also to calm? Cognitive science finds itself in the uncomfortable position of not being able to make the simplest predictions about musical pleasure despite the fact that there are no difficulties in describing the (raw) stimulus. The situation is comparable to the problem of chess-playing before the advent of the concept of heuristic search; there were no hidden variables in the game, yet how to create a competent player remained a mystery.

Current cognitive theories of music concentrate primarily on categorization. For example, Leman (1991) has shown that the cycle of fifths is an emergent property of applying an unsupervised learning algorithm to a representation that is rich in harmonic overtones. A theory of musical categorization, in the absence of hedonic principles, will not yield a theory of musical pleasure, however, for the simple reason that there is no reason to suppose that one category is preferred to any other. It may seem possible to use the results of categorization to graft the drive toward music onto simpler, innate drives. The gap that needs to be bridged here is large, however, and there are few supporting pylons along the way. The pleasure associated with a Bach fugue does not translate well into the theoretical vocabulary of hunger drive reduction, or secondary drive reduction.

Another possible explanation of musical affect has been suggested by Jackendoff (1991) who has shown how a theory of musical parsing may work in conjunction with Meyer's (1956) theory of musical affect. Meyer's governing principle is that affect is generated when a tendency to respond is inhibited. One immediate difficulty with such a theory is that it predicts that a well-known piece, which, by hypothesis, is completely predictable, should not generate any affect. In fact, as Jackendoff points out, good music needs many hearings before being fully appreciated. His solution to this dilemma is to retain Meyer's framework, but claim that the musical parser operates independently of musical memory; thus, the parser continues to have its expectations violated or confirmed regardless of the familiarity of the composition.

Jackendoff's extension is problematic, however, as it rests upon a theory which is weakly predictive of affect. Consider the cadence in Figure 1. Meyer is able to show how the composer may generate affect by evading the cadence, i.e., by postponing the appearance of the tonic (I), thus violating an expectation. What he cannot explain is the ubiquitous presence of this transition in both classical and modern popular music. Assuming it is not evaded, as is often the case, it will be completely expected by anyone familiar with the genre of the

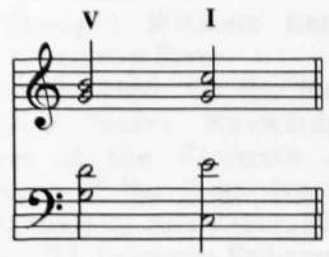

Figure 1. An authentic cadence.

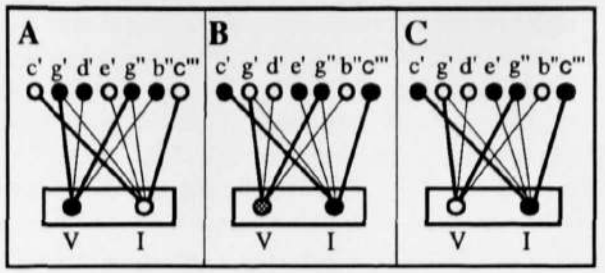

Figure 2. Model of the cadential resolution.

composition. It does not help to transfer generation of affect to the parser, as this processor will also expect the tonic as the most common resolution of the dominant (V). Moreover, predicting the variety in cadential forms, and their relative degree of thrust, places a burden on the expectation theory that it cannot be born easily.

In the next section, it will be shown that a simple connectionist model, in conjunction with an assumption concerning the relation between activation and affect, will suffice to predict the power of the cadence, and variations thereon. This model will then be extended to melodic sequences.

A model of musical resolution

Katz (1993a) has described a model of musical resolution based on a model of resolution in humor (Katz, 1993b). The theory is based on the fact that an activation boost will result when a concept is partially maintained for a short time while a competing concept is triggered. For example, Figure 2 shows a model for the cadential resolution in Figure 1. Panel A illustrates the situation when the network detects the first chord. The appropriate notes are triggered, and activation spreads from these notes to the chord recognition unit via feedforward connections to the chord layer, which is a competitive, winner-take-all subnetwork.

This competitive network will make the unit receiving the most activation fully active (in this case V) and all other chord units fully inactive (in this case, I). Panel B shows what happens when the new chord, I, is detected. The unit for this chord will win the competition, as it now receives the greatest input. The V unit, however, will also be maintained, because of the input from the shared note, g". The effect will be further enhanced if notes at a semitone (such as the b" and c") or whole tone distance (such as the d" and e") are assumed to share an overlapping distributed representation by virtue of their perceived similarity (not shown in the diagram). This explains why notes resolve by step, that is, by falling or rising a whole tone or less. Eventually, I wins the competition, and drives V to a quiescent state, as shown in Panel C.

When a concept is supported while its competitive counterpart is triggered, as in Panel B, the total activation of the competitive layer will be greater than that of the relaxed state, in which only one unit can be active. There are two related reasons why this boost is hypothesized to underlie positive affect. The first derives from the classical aesthetic goal of unity in diversity. To the extent that that the two chords are maintained simultaneously, the network is uniting two concepts that are ordinarily found in opposition. Alternatively one can argue that the perceived coherence of the transitions is proportional to the boost. That is, if the shift from one concept to the next is abrupt, there will be little time when the two are simultaneously active. An activation boost will only result in those cases where there is a smooth, coherent transfer between concepts.

The network dynamics are sufficiently complex such that the theory does not reduce to one of similarity between successive concepts. For example, the theory is able to show why the V to I transition is of greater perceived thrust than the reverse transition of I to V, although the similarity of I to V is the same as V to I. The g", as the root of V in C major, strongly supports V, while as the fifth of I, weakly supports I. Thus, V will receive greater support during transfer from V to I than the I will receive during transfer from I to V, resulting in a larger boost.

Melodic sequences

The goal of this paper is to extend the above model to the appreciation of melodic sequences. Consider the three sequences in figure 3. Each sequence is composed of two phrases, A and B, and in each B resembles A in key aspects. In sequence 1, the first five bars in the melody of Beethoven's Fifth Symphony, the second phrase is identical to the first, but transposed down (diatonically) by a second. In sequence 2, chimes, phrase B repeats the same notes in phrase A but in different order. Sequence 3

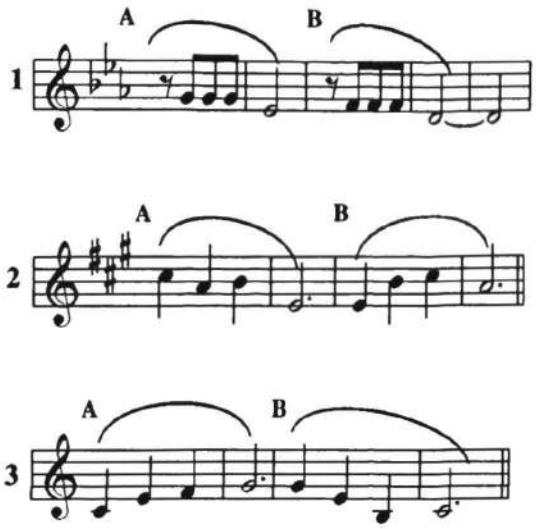

Figure 3. Three melodic sequences.

presents an inexact retrograde, whereby phrase B returns to the tonic note, c, in approximately the same way phrase A approaches the fifth of the key, g.

In each of the three cases, phrase B shares features with A. If a network can be created in which A and B correspond to a unique category, and if the input to this network captures the similarity between A and B, then a larger activation boost will result from the transfer of category A to B than from A to a category that does not resemble A. Thus, a similar model to that described earlier will capture the fact that preferred melodic sequences are those with phrase similarity. Three conditions must be met for this model to work. First, some mechanism must exist to parse the melodic sequence into groups. Next if two groups are heard as similar, this must be reflected in the representation. Finally, there must exist some unsupervised learning mechanism which creates categories corresponding to the groups.

Parsing

A successful parse is necessary for both the understanding and enjoyment of music. The alternative, a sort of continuous categorization that treats all possible groupings on equal footing (as found in Gjerdingen, 1991, e.g.) misses the essence of musical communication as much as any interpretation of verbal utterances that did not have the concepts of the pause and the full stop.

Parsing an intricate piece of music into groups is task rivalled in complexity only by the parsing of an intricate linguistic sentence into parts of speech; accordingly, this section will not offer a complete solution to this problem. Rather, the purpose of this

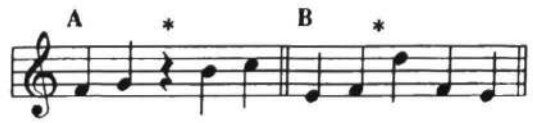

Figure 4. The parsing of simple melodies.

section is to show that two of the preference rules in Lehrdal and Jackendoff's (1983) grouping theory can reduced to a single principle. Their first principle, proximity, is illustrated in case A of Figure 4. The rule states that a group boundary will be preferred at the '*' as the gap between the notes on both sides of the rest is relatively large.

Their other principle, change, is illustrated by melody B in Figure 4. A new group will be preferred when there is a relatively large change in register. If one assumes that the input layer of a network decays in a regular fashion, then proximity is a special case of change. That is, if one compares the current input to the state of a buffering layer, in which the activation of past notes falls off as they recede in time, then a large change will be detected if there is a significant gap between a note's attack and the succeeding note. Thus, a simple rule that detects change (for the purposes of the simulations below, a euclidian measure is used) will be sufficient to detect group boundaries, and reset the categorization process. This parsing mechanism is far from complete, but adequate for the simple melodies in this paper.

Representation

The representation is crafted such that similarity in melodic sequences can be recognized. Figure 5 shows the representation scheme, and the state of the units after the e has been sounded at the end of the first phrase in chimes (melody 2 in Figure 3). The representation is divided into three sets of units.

The first set encodes the pitch and the recency of attack of the note; the strength of a note decays exponentially with time. An alternative is to create a variable window, equal in size to the number of notes in the group. But this would not capture the perceived similarity between notes repeated in different order. In addition, the note fading scheme helps the parsing mechanism.

The second set of units contain interval information. The current scheme, which indicates leap up (a jump of greater than a second), step up, no movement, step down, and leap down, is a compromise between representing the exact set of intervals, and contour information, which would only indicate up, no change, or down. The intervals are stored with a reverbatory circuit, such that, upon

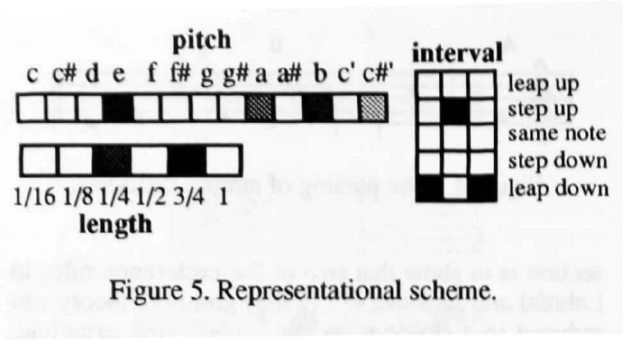

Figure 5. Representational scheme.

receiving the last note of a group of n notes, the interval units will contain in sequential order the n-1 intervals in the group. No neural mechanism of extracting intervals will be proposed here, although there can be no doubt that the auditory cortex somehow extracts this information. Even the musically untrained have little trouble recognizing a transposed (i.e., interval preserved) melody; conversely, musical sophistication is no guarantee of absolute pitch.

Finally, the length units encode the duration of the units, with a similar fade mechanism to the pitch units. In a full representation, it is probably necessary to encode the rhythm in a similar manner to the encoding of the interval information, although the indicated representation will prove adequate for the simulations described below.

Categorization

This section describes an unsupervised learning algorithm such that the strength of the connection between an input unit and a winning unit in the category layer is proportional to the mutual activation between the two units (cf., Rumelhart and Zipser, 1986).

Weights from the input layer to the competitive cluster of units are initially set to 0. At each detected group boundary, the winning category is detected in the following manner. If there is a category unit such that the cosine between the input vector and the weight vector to this unit is greater than a parameter between 0 and 1.0, μ, (cf. Grossberg, 1980), than this is the winner (if there are many such products, the largest is chosen). Otherwise, an uncommitted unit, i.e., a unit which has not yet won a competition, is selected as the winner. Once the winner is found, the weight vector to this unit is set to be the weighted average of the existing weight vector and the current input vector, typically giving greater prominence to the existing vector. The resulting weight vector is then normalized to length 1.0.

The above will suffice for a single layer of recognition, but is deficient if the goal is higher order categorization. Consider, e.g., the common binary melodic form ABA'C, where each of the letters stands for a group of notes. One would like to apply the above algorithm hierarchically to recognize the unity of the first half, AB, with the second, A'C. Hierarchical categorization can be achieved by adding two new layers to the network, one buffering the result of the categorization layer, and one to form higher-order categories. However, if μ is above a critical value, then A and A' will be placed in a different categories, and there will be no similarity between AB and A'C. If μ is too low, then A and A' will be seen as the same, confounding the melody ABAC with ABA'C. One solution to this problem is to form a distributed representation by allowing multiple competitive clusters in each categorization layer, with μ varying from cluster to cluster.

Simulation results

This is the solution adopted for the purposes of this section. The parameter μ ranges from 0.1 to 0.9, corresponding to each of 9 competitive clusters. Each cluster consists of eight units with self-connections set to +0.75, and lateral inhibition to all other units in the cluster set to -0.5. The total activation to each layer is normalized to 1.0 to ensure proper operation of the clusters regardless of the absolute size of the input. A sigmoid transfer function is used, with threshold 0.75. The network consists of four layers, an input layer, a categorization layer, containing the competitive clusters and classifying the input, a layer that buffers the result of this categorization layer, so that the fourth layer can in turn categorize its state.

Testing of all melodies consists of two phases, first a training phase, in which the categories for the melody are acquired, and then a testing phase, in which the mean activation in the second and fourth layers is measured. Training always begins with a an empty network (i.e., all feedforward weights are 0.0); this enables the network to judge the worth of the melody free from interference of familiarity effects.

Hedonic tone and complexity

Intuitively, good music lies somewhere between uniformity and chaos. Vitz (1966) confirmed this by showing that subjects' hedonic tone was an inverted U-shaped curve as the function of the complexity of randomly generated music. Similar results are obtained with the current model. In the graph in Figure 6 each data point is the average of 50 melodies. Mean activation, measuring the number and strength of the activation boosts, is graphed as a function of the range of randomly chosen notes in a 16-note melody (the length of the note is allowed to vary randomly between an eighth and half note for all ranges). Inverted U's are obtained in both the low-order categories, which categorize the parsed groups of

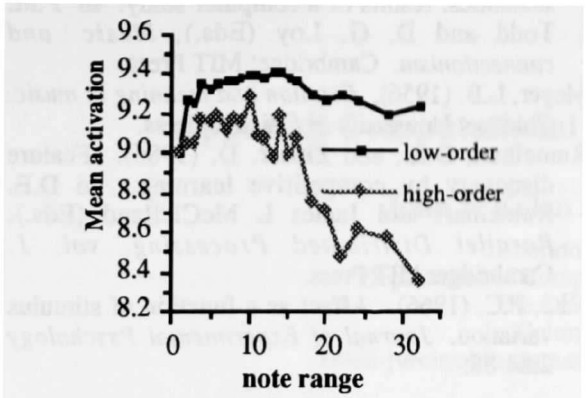

Figure 6. Mean activation of low order and higher order categories as a function of note range.

notes, and high-order categories, which classify the low-order categories. At low complexity, there is high unity between successive groups, but there will be little change in the distributed representation between these groups. Conversely, at high complexity, the distributed representation for successive groups will be almost completely different, but activation boosts associated with these changes will be small, because of the low similarity between the groups. Significantly, the high-order curve is always below that of the low-order, consistent with fact that compositions generated by statistical means tend to lack global unity.

A common melodic form

Melody 3 in Figure 3 was tested in two ways; first in the normal direction, and then by reversing the phrases. An additional assumption was made for this experiment, viz., that notes in the input layer trigger, to a lesser extent, their relative fifth in addition to their fundamental frequency. This assumption can be justified on the basis that all instruments generate overtones; a note and its fifth share the strongest overtones, with the exception of a note and its octave equivalents.

In this experiment, the fifth was given half the value of the fundamental. This assumption yielded the a mean activation in the normal direction of 9.64 and 9.44 in the reverse. Both are above any of the random melodies. The normal direction, however, is preferred for reasons analogous to the preference for the dominant to tonic cadence. Because of the fade mechanism in the representation, the last note of the melody is the most important. In the normal direction, the melody ends on a c, but supports a key component of the category for the first phrase because of the partial triggering of the fifth, or the g. In the reverse direction, the g does not activate the c, but the next highest d, resulting in less support for the category for the first phrase.

A typical song

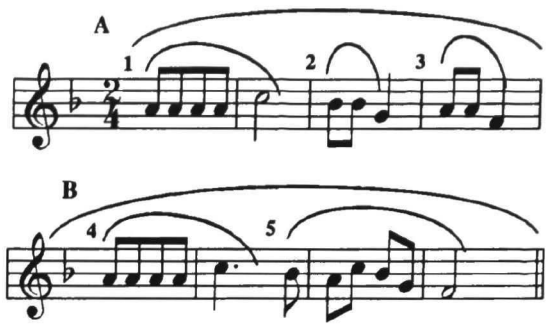

Figure 7. Some folks do.

Figure 7 shows a typical folk song. The network is able to parse the melody into five phrases, because of the relatively long notes that end the phrases. The network also correctly parses these five phrases into higher-order categories; the second and third phrase end up in the same group because of their intervallic and rhythmic similarity. One method of judging musicality is to have people judge the worth of good music that has been altered in some way; they should prefer the closest to the original. A similar experiment has been performed here. A fixed number of notes were altered from their original value to that of a randomly chosen note in a two octave range. The graph in Figure 8 shows mean activation for both the low-order and high-order layers as a function of the number of notes changed. Note that high-order values for zero notes changed (the original melody) are considerably above any of the random melodies; high-order unity is achieved by the near identities of

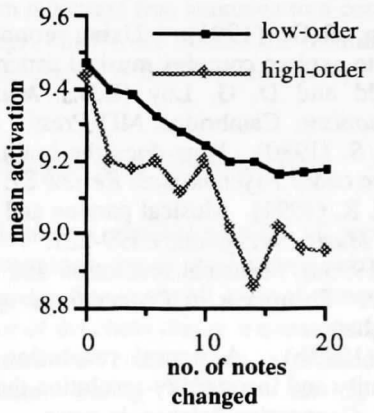

Figure 8. Mean activation for a simple folk song as a function of the degree of resemblance to the original.

phrases 1 and 4. The low-order activation level falls off in a near-linear fashion. The high-order measure is somewhat more sensitive to the disturbance of the melody, reflecting the greater difficulty in achieving global unity.

Discussion

Numerous problems arise in moving from simple folk songs to full-fledged polyphonic music. Here just three corresponding to the three aspects of the model will be mentioned. First, the parsing mechanism will not work with counterpoint, in which an imitative phrase may be begin before the end of the voice it is imitating. Next, the representation only weakly captures an important type of unity in a piece, viz., rhythmic repetition. Finally, the categorization scheme is too efficient to model familiarity effects; by slowing the learning down, it may be possible to show why it takes repeated hearings before a piece is understood and appreciated.

None of these problems are trivial; and yet none fatally detract from the central claim of this paper, that a network model can measure the coherence of the transitions between musical categories by noting the frequency and height of the activation boosts of these transitions. One aspect of musical cognition that is not captured in this model is the possible pleasure associated with affective connotations, learned or otherwise, of a piece of music. This objection could be met by claiming that coherence is a necessary but not sufficient condition for musical pleasure. Alternatively, it may be possible to show that these connotations contribute only minimally to enjoyment, and when present, may be amenable to a theory of the sort proposed in this paper.

References

Gjerdingen, R.O. (1991). Using connectionist models to explore complex musical patterns. In P. M. Todd and D. G. Loy (Eds.), *Music and connectionism.* Cambridge: MIT Press.

Grossberg, S. (1980). How does the brain build a cognitive code? *Psychological Review* 87:1-51.

Jackendoff, R. (1991). Musical parsing and musical Affect. *Music Perception*, 9:199-229.

Katz, B. (1993a). Musical resolution and musical pleasure. To appear in *Proceedings of AISB*, Birmingham.

Katz, B. (1993b). A neural resolution of the incongruity and incongruity-resolution theories of humour. *Connection Science*, in press.

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

Leman, M. (1991). The ontogenesis of tonal semantics: results of a computer study. In P.M. Todd and D. G. Loy (Eds.), *Music and connectionism.* Cambridge: MIT Press.

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

Rumelhart, D.E., and Zipser, D. (1986). Feature discovery by competitive learning. In D.E. Rumelhart and James L McClelland (Eds.), *Parallel Distributed Processing, vol. 1.* Cambridge: MIT Press.

Vitz, P.C. (1966). Affect as a function of stimulus variation. *Journal of Experimental Psychology* 2:84-88.