Introduction

Music teachers often ask their classes to take musical dictation. The teacher commonly plays a short excerpt, perhaps a phrase from a Bach chorale, and then the students attempt to translate the various pitches and durations into the symbols of standard musical notation. Each resulting symbol can be judged true or false, right or wrong. Music teachers also ask their classes to analyze the musical syntax of phrases from Bach chorales. The resulting answers will likely have varying shades of truth or falsity. Students can have legitimate differences of opinion about how to interpret ambiguous musical events or events that have one meaning in prospect and a second in retrospect. And not only students differ in their assessments of complex musical patterns. Acknowledged experts in eighteenth-century music are known to disagree about how, in a given phrase, the chords, rhythms, contours, melodies, textures, timbres, and counterpoint all interact.

Connectionist models offer elegant ways of dealing with multidimensional complexity of the type found in polyphonic, harmonically oriented music. One can, for example, define a musical event as an input pattern of activation that is then transformed by a network of interconnected processing units into an output pattern of activation representing an interpretation of the event. Models that allow for learning often require an explicit teacher prepared to state unambiguously whether, and to what degree, the output pattern is in error (Rumelhart, Hinton, and Williams 1986; Werbos 1988). For relatively low-level musical tasks, like the transcription problem described above, this requirement poses no serious obstacle; but for higher-level tasks, the notion of an explicit teacher becomes problematical. One wonders, for example, if anyone would be comfortable in claiming that one interpretation of a musical phrase is only 69 percent as true as another?

An alternative class of connectionist models comprises the so-called self-organizing networks, networks that learn without explicit teaching inputs (Grossberg 1982; Kohonen 1984). They present an appealing analog of the ordinary listener, someone who without formal training has nevertheless developed a strong sense of how music "works." For many years Stephen Grossberg and his associates have been major contributors to the development of self-organizing networks. In what follows I will explain how Grossberg's adaptive resonance theory (ART) (Grossberg 1976; Carpenter and Grossberg 1987) can be used to develop a self-organized, stable category structure for musical patterns of arbitrary complexity, and I will report on a program entitled L'ART pour l'art (ART for art's sake), which has been taught some early works of Mozart.

An ART Architecture for Analog Input Patterns

Figure 1 shows the basic design of an ART network. Input in the form of a vector of continuous values enters a lower field of processing units (field 1, or $F_1$). There the vector is normalized and contrast-enhanced before proceeding to the upper field of processing units (field 2, or $F_2$). The upper field is a shunting, on-center/off-surround, competitive network, the extreme form of which is sometimes called a winner-take-all design (Carpenter 1989). For any given $F_1$ pattern of activation, the multiplicative connections between $F_1$ and $F_2$ typically give one local $F_2$ population of cells slightly higher excitatory input than is given any other population.
Fig. 1. The basic outline of an adaptive resonance theory (ART) network.

Because of strong lateral inhibition between $F_2$ populations (off-surround interactions) and recurrent excitation within populations (on-center interactions) this dominant population will be able to suppress all or most of the other populations. The surviving population(s) can be viewed as the network's output, as an interpretation of the original input.

Grossberg's networks are intended to emulate principles of human information processing. In the case of the ART network, $F_1$ represents a dynamic, short-term memory (STM) and each $F_2$ population represents a long-term memory (LTM). Short-term memories are simply the activations of cells in $F_1$. Long-term memories result from adjustments to the strengths of $F_1$-$F_2$ connections. These adjustments follow the general form

$$\Delta w_{ij} = Ay_j(x_i - w_{ij}),$$

where $w_{ij}$ is the strength or weight of the connection from cell $i$ in $F_1$ to population $j$ in $F_2$, $y_j$ is the positive activation of population $j$ in $F_2$, $x_i$ is the strength of the excitatory signal from cell $i$ in $F_1$, and $A$ is a constant factor ($A < 1$) determining the rate of learning. As the network is exposed to various input patterns, only winning $F_2$ populations can adjust their connections because only they have a positive term $y_j$. Thus individual $F_2$ populations are adaptive to specific categories of inputs and can eventually match their vector of $w_{ij}$ to the critical feature patterns of a category of input vectors of $x_i$ (Grossberg 1987).

The resonance in ART occurs when learned $F_2$ memories are read out as top-down feedback to $F_1$. This feedback mixes with input patterns of activation within the internal structure of $F_1$, as shown in Fig. 2. $F_1$ is itself a double feedback circuit whose upper and lower loops share the middle of the three internal levels of $F_1$. Normalization of activation patterns occurs during the lateral transmission of signals at each level. This is modeled on the automatic gain-control properties of nonspecific inhibitory interneurons (Carpenter and Grossberg 1987). Contrast-enhancement of patterns and the suppression of noise occur during the vertical transmission of signals between levels. These tasks are accomplished by nonlinear signal functions (the $f(x)$ in Fig. 2), generally of the sigmoid or threshold-linear type. It is the resulting normalized, contrast-enhanced blend of input pattern and $F_2$ feedback that dominates activations in the middle level of $F_1$.

An ART network quickly settles into an $F_1$-$F_2$ feedback resonance, the state during which most learning will occur. What will be learned is not, however, an unmediated input pattern. Rather, the network will learn the above-mentioned mix of the input pattern and an $F_2$ memory. New variations of patterns are automatically interpreted in the context of previously learned pattern prototypes. In this manner the network shields itself against the capricious miscoding of its memories by anomalous input patterns. The resulting hysteresis-like stability can, however, lead to excessive rigidity. In real pattern domains not every new pattern is a variant of previously established pattern types. Some patterns are absolutely and unpredictably new. The ART architecture overcomes this problem with an "orienting subsystem."

As suggested in Fig. 2, there are comparator cells $c_i$ connected to the upper and middle levels of $F_1$. Because of the action of a nonspecific inhibitory interneuron connected to those same levels as well as to the comparator cells (Carpenter and Grossberg 1987), the length, $l$, of the comparator-cell vector $|l = \Sigma c_i^2|^{1/2}$ will vary, assuming a certain range of
inputs, in proportion to the cosine of the activation vectors of the middle and upper levels of $F_1$. In simple terms, this means that as a group, the comparator cells are sensitive to the degree of match between $F_2$ feedback and the mixture of that feedback with the input pattern. When the match is very poor—that is, below a predetermined level—the comparator circuit issues a nonspecific reset signal to $F_2$. That signal suppresses whichever $F_2$ population is currently active. Then, depending on various parameters and the exact nature of the input pattern, either another $F_2$ population with a similar memory will try to meet the comparator criterion or an entirely new population will be selected.

The $F_2$ design of mutually inhibitory populations is, by itself, capable only of what is called competitive learning. “Competitive learning is an essentially nonassociative, statistical learning scheme” (Rumelhart and Zipser 1985) with some similarities to multidimensional scaling or hierarchical clustering analysis. By itself, a competitive learning model may not be able to achieve a stable coding of arbitrarily changing input patterns (Grossberg 1978). But in an ART network, the $F_2$ design does not operate in isolation. $F_2$ and $F_1$ together form a feedback system wherein patterns of input are processed in the context of previously learned pattern categories. The resulting multilevel network can develop a temporally stable coding of an arbitrarily complex and changing domain of patterns (Grossberg 1987). Such a domain is typical of musical patterns.

L’ART pour l’art

Figure 3 shows the conceptual form of an ART network for recognizing and categorizing musical patterns. The lower horizontal disk represents $F_1$, the upper disk $F_2$. The arrows coming from below $F_1$ stand for input sent from feature detectors responding to various aspects of the indicated melody. The three peaks in $F_1$ indicate three cells strongly activated by this input. For the purpose of illustration, let us imagine that these cells correspond directly to the musical concepts do, mi, and sol—the tones of the tonic triad. Joining the two neural fields are heavy vertical lines that represent an excitatory path from every cell in $F_1$ to every population in $F_2$ and vice versa. At the ends of these vertical lines are darkened squares indicating the modifiable strengths of the connections $w_{ij}$ and $w_{ji}$. The lone peak in $F_2$ indicates that a single population has had a winning response to this particular $F_1$ pattern of activation.

The height of each peak in $F_1$ is drawn proportionally to each cell's presumed level of activation. Normalization of patterns across $F_1$ requires that when a new cell receives strong input excitation, the activation levels of previously excited cells
must be decreased. The increasing height of the peaks from do to mi to sol thus does not express the notion that sol causes the intrinsically strongest response. Sol is strongest because it was the tone most recently heard. As the three-note melody would have been performed, each new tone in the present would strongly excite an appropriate cell, which in turn would inhibit the cells still responding to tones from the recent past. When mi sounds, its cell inhibits the previously excited cell for do, and when sol sounds, its cell inhibits the cell for mi and further inhibits the cell for do. This recency effect, as it is called in psychology, permits an ART network to maintain a simple record of temporal order. Indeed, these fading traces of recent events form the ART network’s short-term memory.

By definition short-term memory is limited. As new cells become strongly active, the weakest cells in $F_1$—those retaining traces of the earliest events—become completely inhibited and disappear from short-term memory. For the network to retain a memory of a long series of events, a long melody for instance, it must have a way of remembering individual melodic segments. This is where $F_2$ and possibly higher-level fields $F_3 \ldots F_n$ come into the picture.

If the $F_1$ population in Fig. 3 fully adapts its $F_1$-$F_2$ connections to conform to the $F_2$ pattern of activation, it will develop a long-term memory of an ascending tonic triad. Should that pattern or a slight variant of it reoccur later, the same $F_2$ population will again exhibit a strong response. In fact, because of adaptation to the $F_2$ pattern, later responses will be faster and stronger than earlier ones. Of course the exact $F_2$ pattern that will be learned by the $w_j$ vector of an $F_2$ population depends on how, or whether, the $F_2$ pattern varies. If, for instance, $F_2$ is always presented with an ascending tonic triad performed as, say, three quarter notes, then that exact pattern will be learned by an $F_2$ population. If, on the other hand, rapid, rhythmically irregular ascending tonic triads are randomly interspersed with descending or broken triads, then a more general pattern such as simply “tonic triad” may be learned. Much depends on the entire repertory of patterns to which the network is exposed in its “formative years.”

In an ART network, not only does bottom-up input flow from $F_1$ to $F_2$, but also top-down feedback flows from $F_2$ to $F_1$. We might variously describe this feedback as a template, a prototype, an expectation, or a schema based on the particular $F_2$ population’s experiences with a class of $F_1$ patterns. For example, if $F_2$ cells register the first two tones of an ascending tonic triad, the “ascending-tonic-triad” population in $F_2$ will already begin to respond, and as it does, it will send signals to $F_1$. The top-down feedback to the $F_2$ cells responding to the first two tones of the triad will merely reinforce those cells’ activations. But the feedback to the cell for the third tone of the triad will prime that cell to respond more quickly should the tone actually be heard. The feedback biases the network’s perception of its inputs and makes its responses to its environment more categorical than statistical.

Music is full of surprises, and $F_2$ expectations will not always be born out by developing $F_1$ patterns of activation. In the triadic example just discussed, should quite an unexpected third tone occur, the network’s orienting subsystem will automatically react to the mismatch by suppressing the active $F_2$ population. This action allows a more appropriate population (or a new population) to become maximally active. I suspect that what Leonard Meyer has called the central thesis of the psychological theory of emotions, namely “that emotion or affect is aroused whenever a tendency to respond is arrested or inhibited” (Meyer 1956), may have its physiological basis in the process of a network resetting itself following a mismatch of bottom-up features and a top-down schema.

L’ART pour l’art is a program I have written to test an ART network’s ability to make musically valid categorizations of the type of complex patterns that occur in passages of Mozart. L’ART pour l’art processes information about complex sets of musical features into the form of schematic recognition categories. Input patterns are derived from Mozart’s scores in the following manner. A musical score is viewed as a discrete series of events, each event being the appearance of a new tone in the melody or a new chord. For each event, a tabulation is made of the presence or absence of the 34 specific, low-level musical features listed in Fig. 4. Each
Fig. 4. A listing of individual inputs to $F$, and a symbolic rendering of an $F$, activation vector described in the text.

input vector is mathematically preprocessed to imitate the time-dependent interactions of the new vector with previous $F$ activations. The resulting input vector thus represents the new event as well as fading traces of previous events.

Figure 4 shows a typical set of $F$ activations in symbolic form. The four highest peaks above the field represent the new features of some particular musical event. Old features have lower activations and stand for the inhibited traces of earlier events. It may be instructive to correlate the various peaks with the inputs listed below the field. Starting at the left of the field, notice first a cluster of peaks above the inputs for melodic scale degrees. Presumably the network has detected a descending scale of some type, inasmuch as each scale degree from the third up to the leading tone and tonic has a lower level of activation than the previous degree. The higher tones, occurring first in a descending scale, would have been repeatedly inhibited as each new lower tone is sounded. The presumption of a descending scale is strengthened by the next tall peak to the right, a peak directly above the input for a descending melodic contour.

Further to the right, a second solitary peak stands for the tonic in the bass. Perhaps this is a pedal point, since no other bass tones have registered and there are no activations above inputs for changes in the contour of the bass (inputs 19 and 20). The next group of peaks matches input for scale degrees in one or more inner voices and shows the same descending-scale profile as in the melody, although here the descent is from the fifth to the tonic. Finally, the peak on the far right stands for a contrapuntal dissonance—probably not one in the present, since the peak is not very high, but perhaps one in the recent past.

In Fig. 5, we can see that we have in fact decoded the opening of Mozart's minuet $KV1d$, (transposed here from F to C), a piece written when he was five years old. The exact moment of the network's response is marked by an asterisk at the end of the second measure. Notice that there is a pedal tonic in the bass and the descending lines in the soprano and alto, just as we surmised from Fig. 4. And the presumed recent contrapuntal dissonance turns out to be the passing tones $d''$ and $f''$ over the bass $c$ on.
Fig. 5. L’ART pour l’art shown responding to the point marked with an asterisk at the end of the second measure of five-year-old Mozart’s minuet KV1d (transposed here from F to C). The F₁ activations are a compressed representation of the activations depicted in Fig. 4.

Fig. 6. A musical rendering of the memory—the learned vector of \( w_{ij} \) resident in the dominant F₂ population of Fig. 5. The size of each symbol indicates the relative strength of the memory trace for that feature. Square note-heads stand for inner voices and \( d \) stands for a contrapuntal dissonance. Arrows stand for traces of melodic and bass contour.

the previous beat. Above the musical score is a computer-generated drawing showing the ART network’s response. The lower level \( \{F_1\} \) is the same as in Fig. 4, but with much of the detail compressed. For instance, rather than showing a cluster of separate peaks for melodic scale degrees, the computer drawing shows only the general contour of that cluster. Since short-term memory in an ART network was defined as the vector of active \( F_1 \) cells, in this instance \( F_1 \) demonstrates STM of about two measures of music or perhaps about three seconds—a time typical for human, short-term musical memory [Dowling and Harwood 1986].

As mentioned earlier, long-term memories in an ART network consist of vectors of \( w_{ij} \) modified by individual \( F_2 \) populations. The upper level \( \{F_2\} \) of Fig. 4 shows a high peak toward the middle of the field, where one population has learned to respond strongly to the entire \( F_1 \) pattern. The population developed a memory of that and similar patterns through modifications made to an originally random vector of \( w_{ij} \) over the course of learning trials described below. The resulting vector of \( w_{ij} \) records what Grossberg has called the critical feature pattern [Grossberg 1987], the relative strengths of the musical features associated with this particular \( F_1 \) category of patterns. The critical feature pattern is not, however, simply what was shown in the \( F_1 \) activations of Figs. 3 and 4. The population in question responded to all but the first beat of the two measures shown in Fig. 4. So its critical feature pattern is affected by overlapping experiences of five similar events.

The resulting vector of \( w_{ij} \) has, as would be expected, a very strong trace of the tonic in the bass and no strong trace of either an ascending or a descending bass contour. On the other hand, there is a very strong trace of a descending contour in the melody. The traces for specific scale degrees in the melody and inner voices have a more complex profile affected by metric factors described below. In Fig. 6, I have tried to translate the entire vector of \( w_{ij} \) directly into a type of musical notation so that one can see something of the generic schema that the \( F_2 \) population was able to abstract from its experiences with \( F_1 \) patterns of activation. The relative size of the symbols in Fig. 6 indicate relative strengths of \( w_{ij} \) traces. But the precise weightings of each feature cannot be conveyed in this musical format.

Six of Mozart’s earliest compositions were taught to L’ART pour l’art: KV1a, KV1b, KV1c, KV1d, KV2, and KV3. These pieces are small but nevertheless contain a total of 793 separate musical events. Each learning trial involved teaching all six pieces in varying orders, at slightly different learning rates \((0.01 < \alpha < 0.05)\), and at different tempi \( (F_1 \) interactions vary slightly depending on the rate at which new inputs enter). After 12 trials, a stable category
structure emerged. That is, the same musical event (occurring at the same tempo) would always activate the same \( F_2 \) population.

I arbitrarily allotted \textit{L'ART pour l'art} 25 \( F_2 \) populations. The more \( F_2 \) populations there are, the more distinctions the network can draw. Conversely, the fewer \( F_2 \) populations there are, the more abstract must be each population's memory of an event. For instance, if I had given \( F_2 \) only three populations, it might have been forced to interpret every musical event as one of only three categories—tonic, dominant, or subdominant. While this degree of abstraction is of considerable interest within the discipline of music theory, I do not think it typical of the categorizations made by an ordinary listener. Twenty-five populations seemed a large enough set to promise an interesting level of specificity and yet still small enough to be computationally manageable—an ART network does scale up gracefully to very large systems, but any large parallel network simulated on a serial computer requires an enormous number of calculations.

\( F_2 \) in \textit{L'ART pour l'art} begins inchoate, then slowly organizes itself, by itself, as it is exposed to \( F_1 \) patterns that bear some resemblance to one another. In the early stages of learning, the similarities that the network detects are not always the ones that we would notice first. The contour of the melody and bass, for example, exerted a strong influence on the network's early judgments of similarity, even in cases where, as with dominant and tonic chords, a classical musician might view the patterns as opposites. In truth, it was difficult at times to watch the network making what I considered mistakes. Like a doting parent, I wanted to tell it "No, that's a deceptive cadence!" rather than let it work out the problem for itself, which it eventually did. Conversely, sometimes the network would behave so rationally that I would forget how limited its frame of reference was.

As a case in point, we automatically hear Mozart's predominately two-voice pieces as implying harmonic progressions and the corresponding inner voices. The network, by contrast, had no direct way to obtain such information. In fact, \textit{L'ART pour l'art} was not given knowledge of any a priori correlations among its 34 input features. It did not, for instance, know beforehand that the feature "melodic inflection: sharp" had anything to do with melody or that the feature "bass contour: up" had any correlation with the bass line. When one realizes that \textit{L'ART pour l'art} had to establish the possible correlations among any of 34 features in each of 793 events, and that each event was embedded in a short-term memory carrying forward the decaying traces of 4 or 5 previous events, one gains an appreciation of the difficulty of its task.

With early versions of \textit{L'ART pour l'art} I observed that \( F_2 \) populations seemed to pay too much attention to passing and neighboring tones. This was not just my own prejudice. By categorizing relatively incidental events as unique entities, the network was preventing itself from recognizing the types of underlying similarity that would allow its understanding to progress to more abstract levels. In thinking about this problem, I realized that my perception and the network's perception differed because I automatically heard the pieces in a metric context. Remembering Leonard Meyer's maxim that "in order for meter to exist, some of the pulses in a series must be accented—marked for consciousness—relative to others" (Cooper and Meyer 1960) I realized that the network needed a way to mark an event for consciousness.

One approach would have been to add new inputs to the lower level, inputs that would signal features such as strong beat, weak beat, downbeat, and so forth. I had in fact once tried this approach with a more primitive type of network. The problem is that such inputs may add little to the network's knowledge. The same chord, melody, or progression may occur in many different metric positions, so the network soon learns to ignore the inputs assigned to meter. Strong beat, for example, is not an invariant feature of dominant triad.

A different and I believe more successful approach is to equate marking an event for consciousness with paying more attention to it. By this rationale, a downbeat is strong because we pay strong attention to it. Instead of adding more inputs to the lower level, one can simulate meter by modulating the \( F_1 \) response to input. Input occurring on a downbeat then elicits a larger response than the same input occurring on a weaker beat. I implemented this no-
Fig. 6 for an explanation of the notation. π stands for the trace of a contrapuntal tritone.

Fig. 7. An early Mozart cadence (transposed here from F to C) shown with the eight F₄ memories that it arouses. Please refer to...
A higher-level concept
A sequence of concepts
A concept
A bundle of Features

hierarchy implicit in such a network is not a hierarchy of simple pitches. Here bundles of features are recognized as constituting significant musical events; these are then recognized as occurring in schematic sequences, which may themselves be concatenated into still larger musical schemata. Beyond the very lowest levels of such a network, the units of information are not pitches but various musical ideas, concepts, or schemata.

Conclusion

L’ART pour l’art was designed to test the capabilities of a class of self-organizing, neuronlike networks based on Grossberg’s adaptive resonance theory (ART). The strong claim has been made by Grossberg and others that an ART network could successfully categorize an arbitrarily large and complex domain of analog patterns. The early works of Mozart, while certainly not as large or complex as the works of Mahler or Richard Strauss, still present a pattern domain of considerable variety and sophistication. The fact that L’ART pour l’art, forced to interpret that domain from the perspective of an intentionally spare and insufficient set of input features, was nevertheless able to achieve a musically sound categorization of much of this domain suggests that at least some of the claims for ART architectures may be valid. L’ART pour l’art developed memories of critical feature patterns that resemble not simple chords but the harmonic-contrapuntal complexes referred to by musical theorists as voice-leading combinations. The ability of such a small network to derive categorizations comparable to those employed by musical scholars with expert knowledge of this domain indicates that larger ART networks could prove useful tools in the exploration of complex musical patterns and their myriad interactions.

References