

# The Synergy of Music Theory and AI: Learning Multi-Level Expressive Interpretation

Gerhard Widmer

Department of Medical Cybernetics and Artificial Intelligence, University of Vienna, and  
Austrian Research Institute for Artificial Intelligence,  
Schottengasse 3, A-1010 Vienna, Austria  
gerhard@ai.univie.ac.at

## Abstract

The paper presents interdisciplinary research in the intersection of AI (machine learning) and Art (music). We describe an implemented system that learns expressive interpretation of music pieces from performances by human musicians. The problem, shown to be very difficult in the introduction, is solved by combining insights from music theory with a new machine learning algorithm. Theoretically founded knowledge about music perception is used to transform the original learning problem to a more abstract level where relevant regularities become apparent. Experiments with performances of Chopin waltzes are presented; the results indicate musical understanding and the ability to learn a complex task from very little training data. As the system's domain knowledge is based on two established theories of tonal music, the results also have interesting implications for music theory.

## Introduction

Suppose you were confronted with the following task: you are shown a few diagrams like the one in figure 1, consisting of a sequence of symbols and a graph on top of these which associates a precise numeric value with each symbol. You are then given a new sequence of symbols (see bottom half of fig. 1) and asked to draw the 'correct' corresponding graph, or at least a 'sensible' one. Impossible, you think? Indeed, in this form the problem is extremely hard. It is radically underconstrained, it is not at all clear what the relevant context is (that a single symbol itself does not determine the associated numeric value is clear because the same symbol is associated with different values in fig. 1), and the problem is exacerbated by the fact that the examples are extremely noisy: the same example, if presented twice, will never look exactly the same.

This paper will explain why people are nevertheless capable of solving this problem and will present a computer program that effectively learns this task. The

\*This research was sponsored in part by the Austrian *Fonds zur Förderung der wissenschaftlichen Forschung (FWF)*. Financial support for the Austrian Research Institute for Artificial Intelligence is provided by the Austrian Federal Ministry for Science and Research.

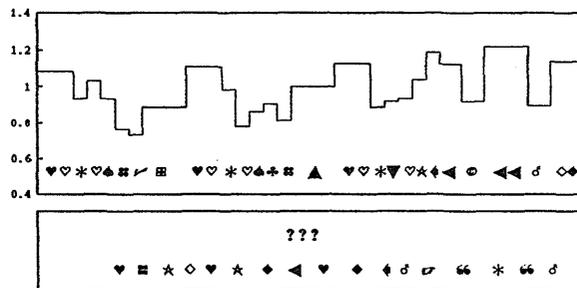


Figure 1: A training example and a new problem.

problem, as the next section will reveal, comes from the domain of tonal music, and it will be solved by combining music-theoretic insights and theories with a hybrid machine learning algorithm. The result is an operational system that learns to solve a complex task from few training examples and produces artistically interesting (if not genuinely original) results. The main points we would like the reader to take home from this are on a general methodological level. This is an interdisciplinary project, and as such it has implications for both AI/machine learning and musicology.

From the point of view of machine learning, the project demonstrates an alternative (though not novel) approach to knowledge-intensive learning. Instead of learning directly from the input data and using the available domain knowledge to guide the induction process, as it is done in many knowledge-based learning systems—e.g., FOCL (Pazzani & Kibler 1992)—we use the domain knowledge (music theory) to restructure and transform the raw input data, to define more abstract target concepts, and to lift the entire problem to a more abstract level where relevant structures and regularities become apparent.

From the point of view of musicology, the interesting result is not only that expressive interpretation can indeed be learned by a machine (at least to a certain degree). The project also indicates that AI and in particular machine learning can provide useful techniques for the empirical validation of general music theories. Our system is based on two well-known

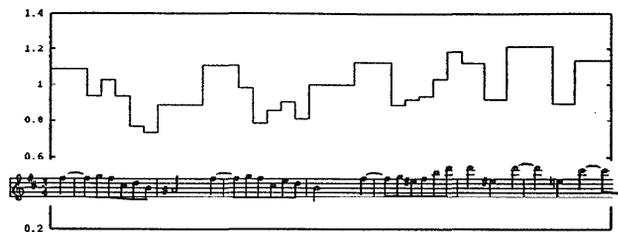


Figure 2: The problem as perceived by a human learner

theories of tonal music (Lerdahl & Jackendoff 1983; Narmour 1977), and an analysis of the learning results provides empirical evidence for the relevance and adequacy of the constructs postulated by these theories.

### A closer look at the problem

To return to the abstract problem in the previous section, why is it that people are able to tackle it successfully? There are two simple reasons: (1) the problem is presented to them in a different form, and (2) they possess a lot of knowledge that they bring to bear on the learning task (mostly unconsciously). To unveil the secret, the people learning this task are music students learning to play some instrument, and to them the problem presents itself roughly as shown in fig. 2. The meaningless symbols from fig. 1 are now the notes of a melody (incidentally, the beginning of Chopin's Waltz Op.69 no.2), and the graph on top plots the relative *loudness* with which each note has been played by a performer. What students learn from such examples is general principles of *expressive performance*: they learn to play pieces of music in an expressive way by continuously varying loudness or tempo, and they learn that by looking at the score as written and simultaneously listening to real performances of the piece. That is, the graph is *heard* rather than seen.

Generally, expressive interpretation is the art of 'shaping' a piece of music by varying certain musical parameters during playing, e.g., speeding up or slowing down, growing louder or softer, placing micro-pauses between events, etc. In this project, we concentrate on the two most important expression dimensions, *dynamics* (variations of loudness) and *rubato* (variations of local tempo). The relevant musical terms are *crescendo* vs. *diminuendo* (increase vs. decrease in loudness) and *accelerando* vs. *ritardando* (speeding up vs. slowing down), respectively. Our program will be shown the melodies of pieces as written and recordings of these melodies as played expressively by a human pianist. From that it will have to learn general principles of expressive interpretation.

Why should the learning problem be easier when presented in the form of fig. 2 rather than fig. 1? The difference between the two representations is that the latter offers us an *interpretation framework* for the symbols; we recognize notes, we recognize patterns (e.g., measures, ascending or descending lines, etc.),

we know that the note symbols encode attributes like duration, tone height, etc. When listening to the piece, we hear more than just single, unrelated notes—we hear the rhythmic beat, we hear groups that belong together, we hear melodic, rhythmic, and other patterns, and we associate the rise and fall of loudness with these groups and patterns. In short, we have additional *knowledge* about the task, which helps us to *interpret* the input.

Our learning program will also need such knowledge if it is to effectively learn expressive interpretation from examples. Music theory can tell us more precisely what the relevant knowledge might be.

### What music theory tells us

Expressive performance has only fairly recently become a topic of central interest for cognitive musicology. There is no general theory of expression, but two assumptions are widely agreed upon among theorists, and these form the basis of our approach:

1. Expression is not arbitrary, but highly correlated with the *structure* of music as it is perceived by performers and listeners. In fact, expression is a means for the performer to emphasize certain structures and maybe de-emphasize others, thus conducting the listener to 'hearing' the piece as the performer understands it.
2. Expression is a *multi-level* phenomenon. More precisely, musical structure can be perceived at various levels, local and global, and each such structure may require or be associated with its own expressive shape. Structures and expressive shapes may be nested hierarchically, but they can also overlap, reinforce each other, or conflict.

The notion of *musical structure* is fundamental. It is a fact that listeners do not perceive a presented piece of music as a simple sequence of unrelated events, but that they immediately and automatically interpret it in structural terms. For instance, they segment the flow of events into 'chunks' (motives, groups, phrases, etc.); they intuitively hear the *metrical structure* of the music, i.e., identify a regular alternation of strong and weak beats and know where to tap their foot. Linearly ascending or descending melodic lines are often heard as one group, and so are typical rhythmic figures and other combinations of notes. Many more structural dimensions can be identified, and it has been shown that acculturated listeners extract these structures in a highly consistent manner, and mostly without being aware of it. This is the (unconscious) musical 'knowledge' that listeners and musicians automatically bring to bear when listening to or playing a piece.

What music theory tells us, then, is that the level of individual notes is not adequate, neither for understanding expressive performances, nor for learning. Analyzing an expressive performance without structural understanding would mean trying to make sense

of figure 1 without being able to interpret the symbols. Expression decisions are not a function of single notes, but usually refer to larger-scale structures (e.g., ‘emphasize this phrase by slowing down towards the end’). That is the level on which the decision rules should be represented; it is also the level on which musicians would discuss a performance.

The design of our system has been guided by these insights. We have selected two well-known theories of tonal music—Lerdahl & Jackendoff’s (1983) *Generative Theory of Tonal Music* and Narmour’s (1977) *Implication-Realization Model*—as the conceptual basis. Both theories postulate certain types of structures that are claimed to be perceivable by human listeners. These types of structures provide the abstract vocabulary with which the system will describe the music. As the structures are of widely varying scope—some consist of a few notes only, others may span several measures—and as expressive patterns will be linked to musical structures, the system will learn to recognize and apply expression at multiple levels.

### From theoretical insights to a strategy

The *raw training examples* as they are presented to the system consist of a sequence of notes (the melody of a piece) with associated numeric values that specify the exact loudness and tempo (actual vs. notated duration), respectively, applied to each note by the performer. However, as observed above, the note level is not adequate. We have thus implemented a *transformation strategy*. The system is equipped with a preprocessing component that embodies its knowledge about structural music perception. It takes the raw training examples and transforms them into a more abstract representation that expresses roughly the types of structures human listeners might hear in the music. In this step also the target concepts for the learner are transformed to the appropriate level of granularity by identifying relevant chunks and associating them with higher-level patterns in the expression (dynamics and tempo) curves. Learning then proceeds at this abstraction level, and the resulting expression rules are also formulated at the structure level. Likewise, when given a new piece to play, the system will first analyze it and transform it into an abstract form and then apply the learned rules to produce an expressive interpretation.

### Transforming the problem

The problem transformation step proceeds in two stages. The system first performs a musical analysis of the given melody. A set of analysis routines, based on selected parts of the theories by Lerdahl and Jackendoff (1983) and Narmour (1977), identifies various structures in the melody that might be heard as units or chunks by a listener or musician. The result is a rich annotation of the melody with identified structures. Fig. 3 exemplifies the result of this step with

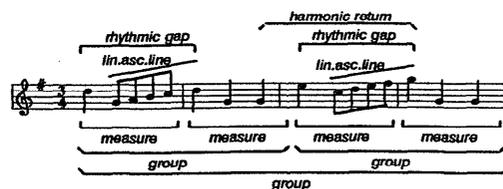


Figure 3: Structural interpretation of part of minuet.

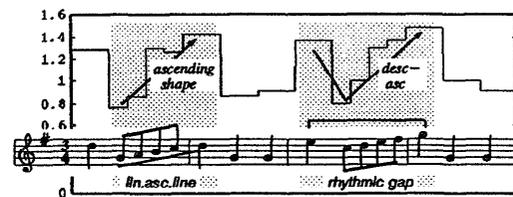


Figure 4: Two of the expressive shapes found.

an excerpt from a simple Bach minuet. The perceptual chunks identified here are four *measures* heard as rhythmic units, three *groups* heard as melodic units or ‘phrases’ on two different levels, two *linearly ascending melodic lines*, two rhythmic patterns called *rhythmic gap fills* (a concept derived from Narmour’s theory), and a large-scale pattern labelled *harmonic departure and return*, which essentially marks the points where the melody moves from a stable to a less stable harmony and back again. It is evident from this example that the structures are of different scope, some completely contained within others, some overlapping.

In the second step, the relevant abstract target concepts for the learner are identified. The system tries to find prototypical *shapes* in the given expression curves (dynamics and tempo) that can be associated with these structures. Prototypical shapes are rough trends that can be identified in the curve. The system distinguishes five kinds of shapes: *even\_level* (no recognizable rising or falling tendency of the curve in the time span covered by the structure), *ascending* (an ascending tendency from the beginning to the end of the time span), *descending*, *asc.desc* (first ascending up to a certain point, then descending), and *desc.asc*. The system selects those shapes that minimize the deviation between the actual curve and an idealized shape defined by straight lines. The result of this analysis step are pairs  $\langle \text{musical structure, expressive shape} \rangle$  that will be given to the learner as training examples.

Fig. 4 illustrates this step for the dynamics curve associated with the Bach example (derived from a performance by the author). We look at two of the structures found in fig. 3: the ascending melodic line in measures 1–2 has been associated with the shape *ascending*, as the curve shows a clear ascending (*crescendo*) tendency in this part of the recording. And the ‘rhythmic gap fill’ pattern in measures 3–4 has been played with a *desc.asc* (*decrescendo – crescendo*) shape.

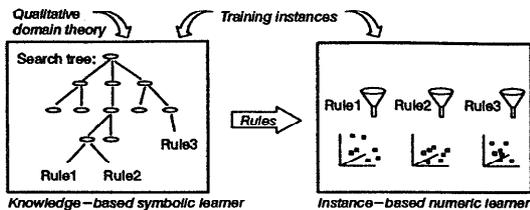


Figure 5: Schema of learning algorithm IBL-SMART.

### Learning expression rules: IBL-SMART

The results of the transformation phase are passed on to a learning component. Each pair *<musical structure, expressive shape>* is a training example. Each such example is further described by a quantitative characterization of the shape (the precise loudness/tempo values (relative to the average loudness and tempo of the piece) of the curve at the extreme points of the shape) and a description, in terms of music-theoretic features, of the structure and the notes at its extreme points (e.g., note duration, harmonic function, metrical strength, ...). Some of these descriptors are symbolic (nominal), others numeric.

In abstract terms, the problem is then to learn a numeric function: given the description of a musical structure in terms of symbolic and numeric features, the learned rules must decide (1) which shape to apply and (2) the precise numeric dimensions of the shape (e.g., at which loudness level to start, say, a crescendo line, and at which level to end it).

The learning algorithm used in our system is IBL-SMART (Widmer 1993). IBL-SMART is a multistrategy learner in two respects: at the top level, it integrates symbolic and numeric learning; and the symbolic component integrates various plausible reasoning strategies so that it can utilize a given *domain theory* (possibly incomplete and imprecise/qualitative) to bias the induction process.

The second aspect is not relevant here, as we have no explicit domain theory—the musical knowledge is used in the preprocessing stage. The integration of symbolic and numeric learning is what is required here, and that is realized in a quite straightforward way in IBL-SMART: the program consists of two components (see fig. 5), a symbolic rule learner and an instance-based numeric learner. The symbolic component is a non-incremental algorithm that learns DNF rules by growing an explicit discrimination or refinement tree in a top-down fashion. The basic search strategy is inspired by the ML-SMART framework (Bergadano & Giordana 1988): a best-first search, guided by coverage and simplicity criteria, is conducted until a set of hypotheses is found that covers a sufficient number of positive examples. In our case, the *target concepts* for the symbolic learner are the different expressive shapes, i.e., it learns to determine the appropriate general shape to be applied to a musical structure.

The numeric component of IBL-SMART is an

instance-based learner that in effect builds up *numeric interpolation tables* for each learned symbolic rule to predict precise numeric values. It stores the instances with their numeric attribute values and can predict the target values for some new situation by numeric interpolation over known instances. The connection between these two components is as follows: each rule (conjunctive hypothesis) learned by the symbolic component describes a subset of the instances; these are assumed to represent one particular subtype of the concept to be learned. All the instances covered by a rule are given to the instance-based learner to be stored together in a separate instance space. Predicting the target value for some new situation then involves matching the situation against the symbolic rules and using only those numeric instance spaces for prediction whose associated rules are satisfied. The symbolic learner effectively partitions the space for the instance-based method, which then constructs highly specialized numeric predictors. The basic idea is somewhat reminiscent of the concept of *regression trees* (Breiman *et al.* 1984). For a more detailed presentation of the algorithm, the reader is referred to (Widmer 1993).

### Applying learned rules to new problems

When given the score of a new piece (melody) to play expressively, the system again first transforms it to the abstract structural level by performing its musical analysis. For each of the musical structures found, the learned rules are consulted to suggest an appropriate expressive shape (for dynamics and rubato). The interpolation tables associated with the matching rules are used to compute the precise numeric details of the shape. Starting from an even shape for the entire piece (i.e., equal loudness and tempo for all notes), expressive shapes are applied to the piece in sorted order, from shortest to longest. That is, expression patterns associated with small, local structures are applied first, and more global forms are overlaid later. Expressive shapes are overlaid over already applied ones by averaging the respective dynamics and rubato values. The result is an expressive interpretation of the piece that pays equal regard to local and global expression patterns, thus combining micro- and macro-structures.

## Experimental Results

This section briefly presents some results achieved with waltzes by Frédéric Chopin. The training pieces were five rather short excerpts (about 20 measures each) from the three waltzes Op.64 no.2, Op.69 no.2 (see fig.2), and Op.70 no.3, played by the author on an electronic piano and recorded via MIDI. The results of learning were then tested by having the system play other excerpts from Chopin waltzes. Here, we can only show the results in graphic form.

As an example, fig. 6 shows the system's performance of the beginning of the waltz Op.18 after learning from the five training pieces. The plots show

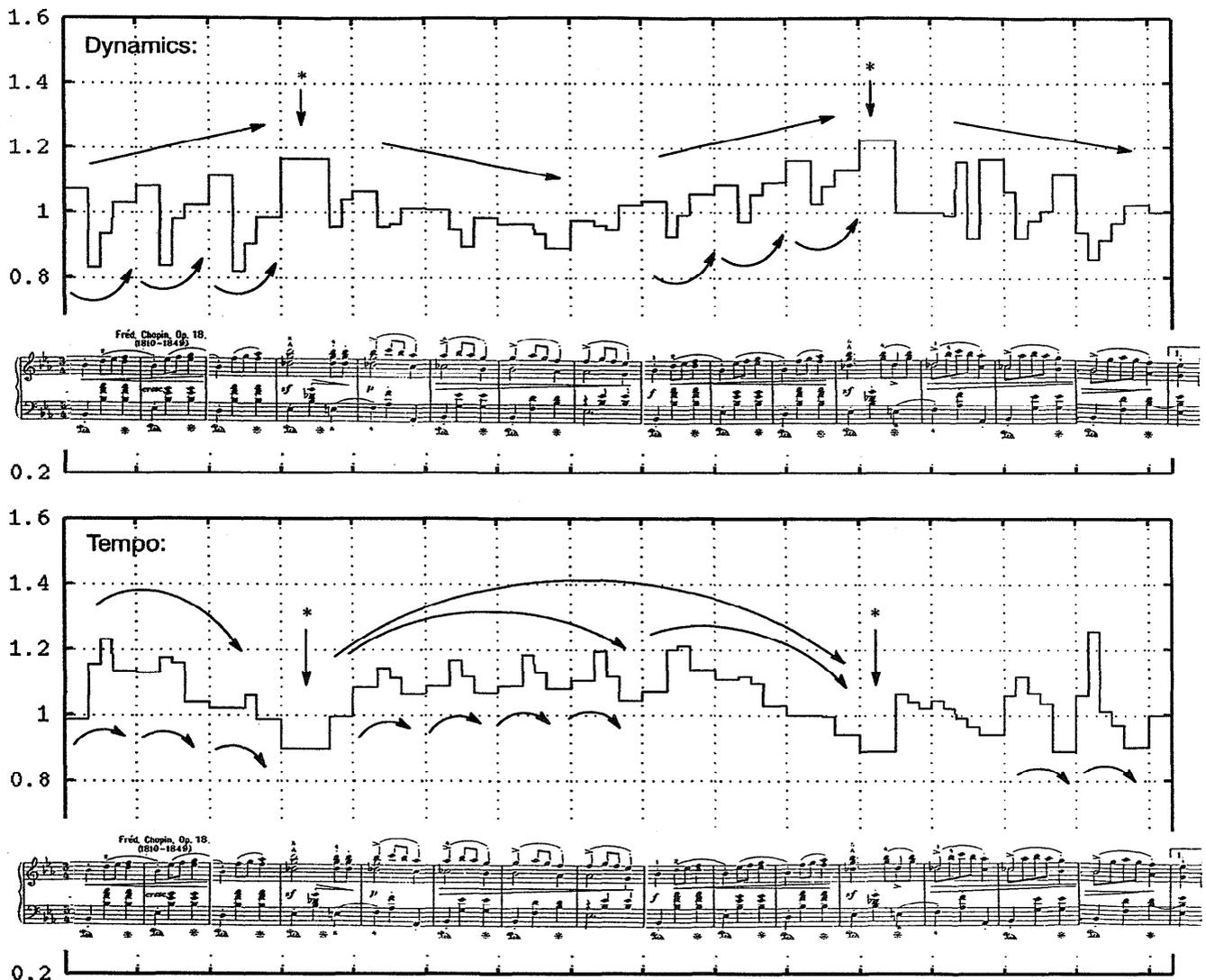


Figure 6: Chopin Waltz op.18, Eb major, as played by learner: *dynamics* (top) and *tempo* (bottom).

the loudness (*dynamics*) and tempo variations, respectively. A value of 1.0 means average loudness or tempo, higher values mean that a note has been played louder or faster, respectively. The arrows have been added by the author to indicate various structural regularities in the performance. Note that while the written musical score contains some explicit expression marks added by the composer (or editor) — e.g., commands like *cresc*, *sf* or *p* and graphical symbols calling for large-scale crescendo and decrescendo — the system was not aware of these; it was given the notes only.

It is difficult to analyze the results in a quantitative way. One could compare the system's performance of a piece with a human performance of the same piece and somehow measure the difference between the two curves. However, the results would be rather meaning-

less. For one thing, there is no single correct way of playing a piece. And second, relative errors or deviations cannot simply be added: some notes and structures are more important than others, and thus errors are more or less grave.

In a qualitative analysis, the results look and sound musically convincing. The graphs suggest a clear understanding of musical structure and a sensible shaping of these structures, both at micro and macro levels. At the macro level (arrows above the graphs), for instance, both the dynamics and the tempo curve mirror the four-phrase structure of the piece. In the dynamics dimension, the first and third phrase are played with a recognizable crescendo culminating at the end point of the phrases (the B $\flat$  at the beginning of measures 4 and 12). In the tempo dimension, phrases (at least the first

three) are shaped by giving them a roughly parabolic shape—speeding up at the beginning, slowing down towards the end. This agrees well with theories of rubato published in the music literature (Todd 1989).

At lower levels, the most obvious phenomenon is the phrasing of the individual measures, which creates the distinct waltz 'feel': in the dynamics dimension, the first and metrically strongest note of each measure is emphasized in almost all cases by playing it louder than the rest of the measure, and additional melodic considerations (like rising or falling melodic lines) determine the fine structure of each measure. In the tempo dimension, measures are shaped by playing the first note slightly longer than the following ones and then again slowing down towards the end of the measure.

The most striking aspect is the close correspondence between the system's variations and Chopin's explicit marks in the score (which were not visible to the system!). The reader trained in reading music notation may appreciate how the system's dynamics curve closely parallels Chopin's various crescendo and decrescendo markings and also the *p* (*piano*) command in measure 5. Two notes were deemed particularly worthy of stress by Chopin and were explicitly annotated with *sf* (*sforzato*): the Bb's at the beginning of the fourth and twelfth measures. Elegantly enough, our program came to the same conclusion and emphasized them most extremely by playing them louder and longer than any other note in the piece; the corresponding places are marked by arrows with asterisks in fig. 6.

Experiments with other Chopin waltzes produced results of similar quality. Preliminary results with songs by Franz Schubert are also encouraging, but suggest that an overabundance of musical structures might degrade the quality somewhat. This indicates the need for a more refined shape combining strategy.

## Summary and Discussion

This paper has presented a system that learns to solve a complex musical task from a surprisingly small set of examples and produces artistically interesting results. The essence of the method is (1) a theory-based transformation of the learning problem to an appropriate abstraction level and (2) a hybrid symbolic/numeric learning algorithm that learns both symbolic decision rules and predictors of precise numeric values.

What really made the problem solvable—and this is the main point we would like to make—is the interdisciplinary and principled approach: combining machine learning techniques with a solid analysis of the task domain and using existing theories of the domain as a sound basis. The result is a system that is of interest to both fields involved, machine learning and music.

From the point of view of machine learning, using available domain knowledge to transform the learning problem to an abstraction level that makes hidden regularities visible is a viable alternative to more 'standard' knowledge-based learning, where learning pro-

ceeds at the level of the original data, and the knowledge is used to bias induction towards plausible generalizations. This approach has also been advocated by a number of other researchers, most notably (Flann & Dietterich 1989). That does not preclude the additional use of domain knowledge for guiding the induction process. Indeed, though the performances produced by our system are musically sensible, the rules it constructs do not always correspond to our musical intuition. To further guide the system towards interpretable rules we plan to supply it with a partial *domain theory* that specifies relevant dependencies between various domain parameters. This will require no changes to the system itself, because IBL-SMART is capable of effectively taking advantage of incomplete and imprecise domain theories (Widmer 1993).

For musicology, the project is of interest because its results lend empirical support to two quite recent general theories of tonal music. In particular, the role of Narmour's music theory is strengthened by our results. Some music researchers claim that grouping (phrase) structure is the essential carrier of information for expressive phrasing. An analysis of the results of our system, however, suggests that melodic surface patterns derived from Narmour's theory are equally important and determine or explain to a large extent the microstructure of expression. We would generally propose our methodology (using established artistic or other theories as a basis for programs that learn from real data) as a fruitful empirical validation strategy.

## References

- Bergadano, F., and Giordana, A. 1988. A knowledge intensive approach to concept induction. In *Proceedings of the Fifth International Conference on Machine Learning*. Ann Arbor, MI.
- Breiman, L.; Friedman, J.; Olshen, R.; and C. Stone, C. 1984. *Classification and Regression Trees*. Belmont, CA: Wadsworth.
- Flann, N., and Dietterich, T. 1989. A study of explanation-based methods for inductive learning. *Machine Learning* 4(2):187–226.
- Lerdahl, F., and Jackendoff, R. 1983. *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.
- Narmour, E. 1977. *Beyond Schenkerism*. Chicago University Press.
- Pazzani, M., and Kibler, D. 1992. The utility of knowledge in inductive learning. *Machine Learning* 9(1):57–94.
- Todd, N. 1989. Towards a cognitive theory of expression: The performance and perception of rubato. *Contemporary Music Review* 4:405–416.
- Widmer, G. 1993. Plausible explanations and instance-based learning in mixed symbolic/numeric domains. In *Proceedings of the 2nd Intl. Workshop on Multistrategy Learning*. Harper's Ferry, W.VA.