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
Tel: +43 - 1 - 53532810
Fax: +43 - 1 - 5320652
e-mail: gerhard@ai.univie.ac.at

Content areas: Machine Learning, Music, Art, Perception

Extended version of a paper appearing in *Proceedings of the 11th National Conference on Artificial Intelligence (AAAI-94)*, Seattle, WA, August 1994.

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

Content areas: Machine Learning, Music, Art, Perception

Abstract

The paper presents genuinely 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. This 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.

1 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 figure 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 figure 1), and the problem is exacerbated further 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 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 new, 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 paper will describe the general approach and the methods used and will present some experimental 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.

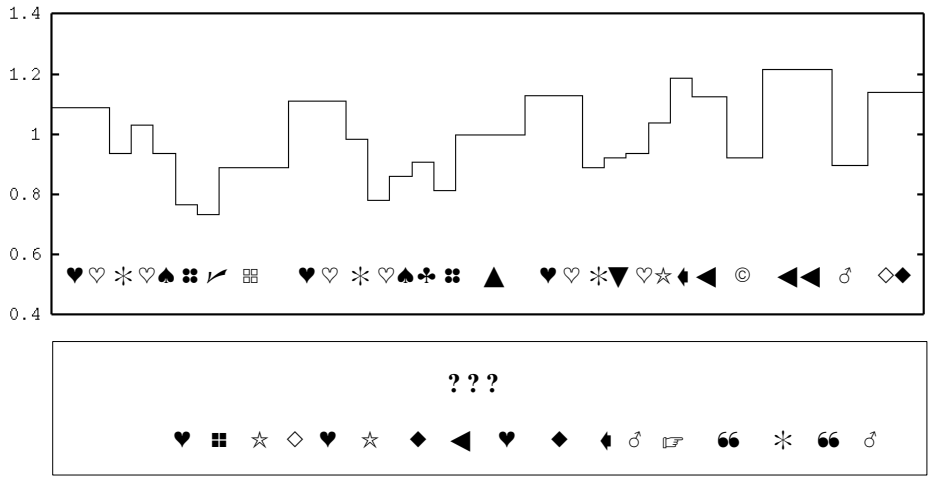

Figure 1: An example to learn from, and a new problem.

From the point of view of machine learning, the project demonstrates an alternative 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 most knowledge-based learning systems—e.g., ML-SMART (Bergadano & Giordana, 1988) or 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. That is also the level on which musicians tend to discuss the problem.

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

And finally, for music in general these results suggest the possibility of building flexible interactive musical instruments that could adapt to a human performer's style of playing; that will be of interest to composers and performers of electronic music.

2 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

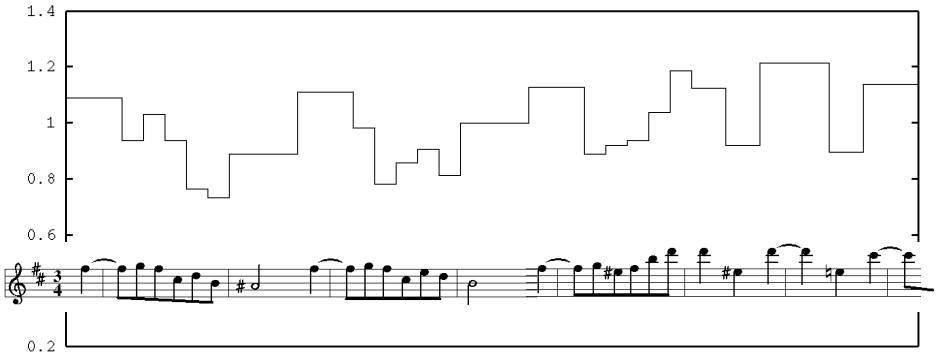

Figure 2: The problem as perceived by a human learner (*dynamics* curve).

roughly as shown in figure 2. The meaningless symbols from figure 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—and what our program is going to learn—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 figure 2 rather than figure 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, and how it is related to musical expression.

3 What music theory tells us about the problem

Expressive performance has only fairly recently become a topic of central interest for cognitively oriented music research. 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 conducing the listener to 'hearing' the piece as the performer understands it.[1]

2. Expression is a *multi-level* phenomenon (Clarke, 1987; Sloboda, 1985). More precisely, musical structure can be perceived at various levels, local and global, in a piece of music, 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 such 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. To enable our learning system to 'make sense' of its input, we will have to equip it with knowledge about musical structure.

[1] A clarifying remark to readers who feel that we are trivializing the artistic phenomenon of expressive performance by reducing it to a function of structural patterns in the music: We are not talking here about the highly artistic details that distinguish a great pianist or other performer, and that derive in part from his/her deep understanding of music history, experience with styles, social circumstances, and artistic intentions. What is being investigated here is the "rational" component of expression, the types of musical behavior and understanding that are more or less common and agreed upon among musicians—in other words, what a music student must learn in order to produce acceptable performances.

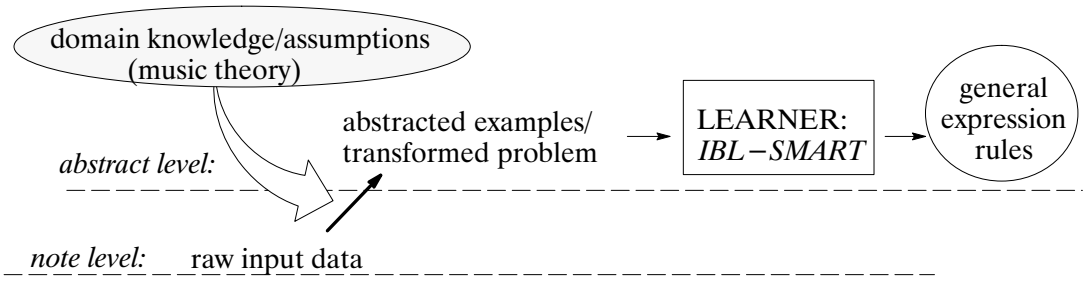

Figure 3: Schema of general strategy.

If our approach is to be musically plausible, and if it is to produce results that are of interest to music theory, two requirements must be met. First, the musical concepts and background knowledge to be supplied to the system must be carefully selected; they should correspond to cognitively plausible concepts that we may assume are shared by most music listeners (from our part of the world). And second, the system must be able to learn and apply expression knowledge at multiple structural levels.

To ensure this, our system has been built on solid music-theoretic grounds. We have selected two well-known theories of tonal music—Lerdahl and Jackendoff's (1983) *Generative Theory of Tonal Music* and Narmour's (1977) *Implication-Realization Model*—as the conceptual basis. Both of these theories postulate certains 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 these structures are of widely varying scope—some consist of two or three 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.

4 Translating theoretical insights into 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 numeric values associated with each note that specify the exact loudness and tempo (actual vs. notated duration) applied to the note by the performer. However, as observed above, the note level is not adequate. We have thus implemented a *transformation strategy* (see figure 3). 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 it to produce an expressive interpretation.

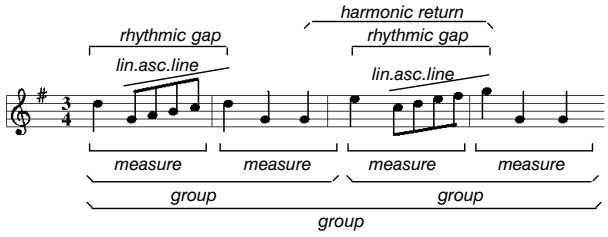

Figure 4: Structural interpretation of part of Bach minuet.

4.1 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. Figure 4 exemplifies the result of this step with 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 these structures are of different scope, some hierarchically 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` (first descending, then ascending). 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 <*musical structure, expressive shape*> that will be passed to the learner as training examples.

Figure 5 illustrates this step for the dynamics curve associated with the Bach example (derived from a performance by the author). We take a look at two of the structures found in figure 4: 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.

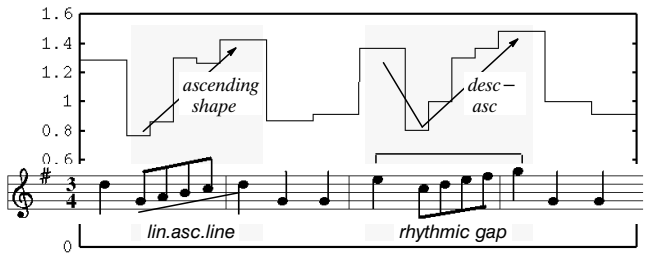

Figure 5: Two of the expressive shapes found in Bach recording.

4.2 Learning qualitative/quantitative rules

The results of the transformation phase are passed on to a learning component. Each pair <*musical structure, expressive shape*> is a training example; more precisely, each such example is characterized by

- the type of structure,
- the type of expressive shape applied to it by the performer,
- 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),
- 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, ...).

The desired *output* of the learning component is a set of general rules that decide, given the description of a musical structure, what kind of expressive shape should be applied to it, and exactly *how much* crescendo, accelerando, etc. should be applied.

The *learning component* itself is based on a new, specially designed learning algorithm named IBL-SMART. In abstract terms, the problem is 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). Standard machine learning algorithms are not usable here. The algorithm IBL-SMART basically integrates a symbolic and a numeric generalization strategy. The symbolic component learns explicit rules that determine the appropriate shape for a musical structure, and the numeric part is an instance-based learner (Aha et al., 1991) that in effect builds up numeric interpolation tables for each learned symbolic rule to predict precise numeric values. 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). The details of the algorithm cannot be discussed here, the reader is referred to (Widmer, 1993) for a detailed presentation.

In any event, the output of the learning component is a set of symbolic decision rules, each associated with numeric interpolation tables. The rules apply rough expressive shapes

to musical structures in some new piece, and the interpolation tables determine the exact expression values to be applied.

4.3 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 overlayed later. Expressive shapes are overlayed 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. The resulting interpretation can then be played via MIDI on an electronic piano.

5 Experimental Results

A number of experiments with different musical styles—from simple Bach minuets all the way to jazz pieces from the swing and bebop eras—were performed. Here we will briefly show some results achieved with waltzes by Frédéric Chopin. The training pieces were five rather short excerpts (about 20 measures on average) from the three waltzes Op.64 no.2, Op.69 no.2 (see figure 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. As we cannot attach recordings to this paper, we will present the results in graphic form (actual recordings will be played at the conference).

As an example, figures 6 and 7 show the system's performance of the beginning of the waltz Op.18 after learning from the five training pieces. The figures plot 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 (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 measure the average difference between the two curves, or determine the percentage of agreement in the number of notes that are played with crescendo and diminuendo, say. However,

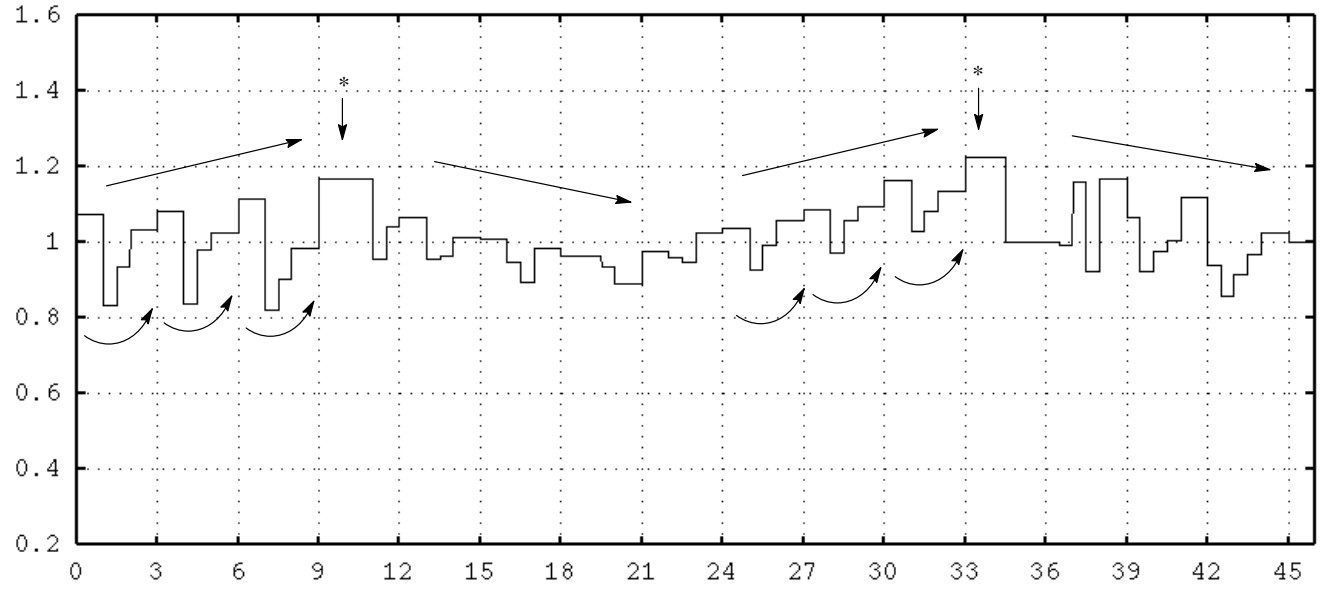

Figure 6: Chopin Waltz op.18, E♭ major, as played by **learner** *(dynamics)*.

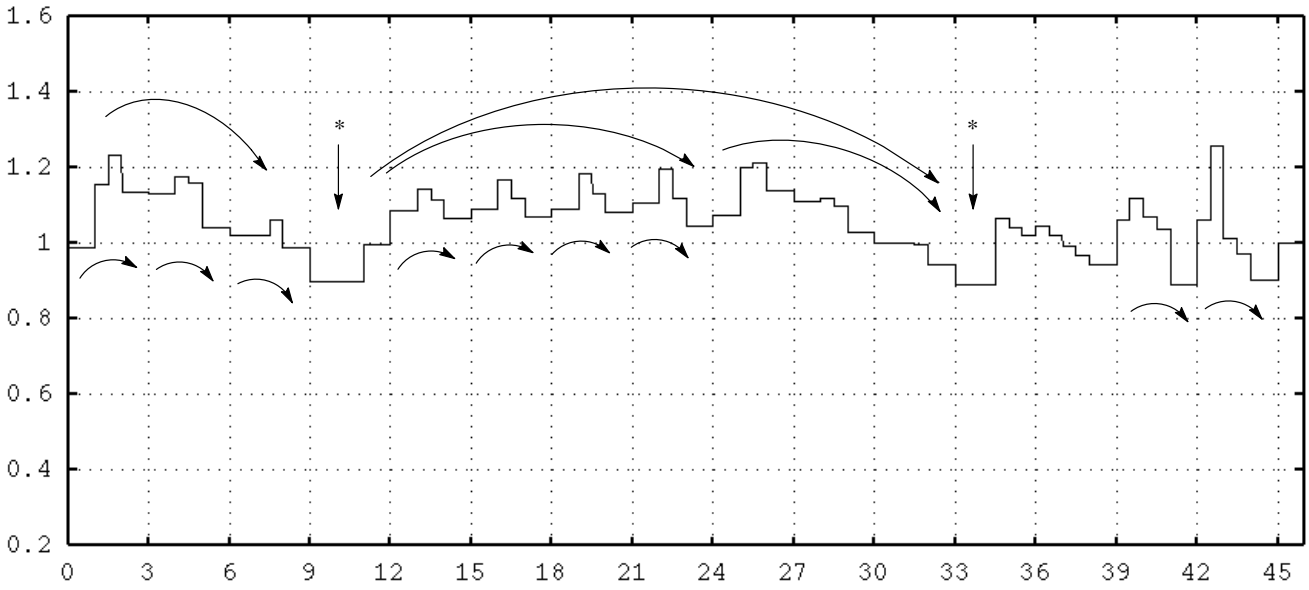

Figure 7: Chopin Waltz op.18, E♭ major, as played by **learner** *(tempo)*.

the results would be rather meaningless. For one thing, there is no single correct way of playing a piece. Also, 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. And third, the multi-level behavior is important, and again, that is difficult to quantify.

In a qualitative analysis, the results look and sound musically convincing. The graphs suggest a clear understanding of musical structure and a musically 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♭ at the beginning of the fourth and twelfth measures—see positions (beats) 9 and 33 in the plot). 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 (e.g., Todd, 1989).

At lower structural levels, the most obvious phenomenon is the phrasing of the individual measures, which creates the distinct waltz 'feel': in the dynamics dimension (figure 6), 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 (figure 7), measures are shaped by playing the first note slightly longer than the following ones (i.e., extending its duration relative to the following notes) 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 B♭'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 figs. 6 and 7.

Just for comparison, figure 8 shows the dynamics curve from an independent recording of the same piece by the author. There are strong similarities at the macro level. However, the author's own performance is embarrassingly poor: it is much less regular and controlled in the fine details (due to the poor keyboard of the electronic piano and the author's far from perfect piano technique). Note that the training pieces from which the system learned were of no better quality. That the system learns to produce smooth performances from bad examples is in part due to the abstraction of *expressive shapes* (see section 4.1) from the low-level details of an example performance.

Figure 9, finally, gives another indication of the system's musical competence by showing the tempo curve of the program's performance of the second part of the waltz Op.64 no.2. Again, note the G at the beginning of measure 7, explicitly marked for emphasis by a > mark in the score, and the way the system stresses the note with an extreme ritardando.

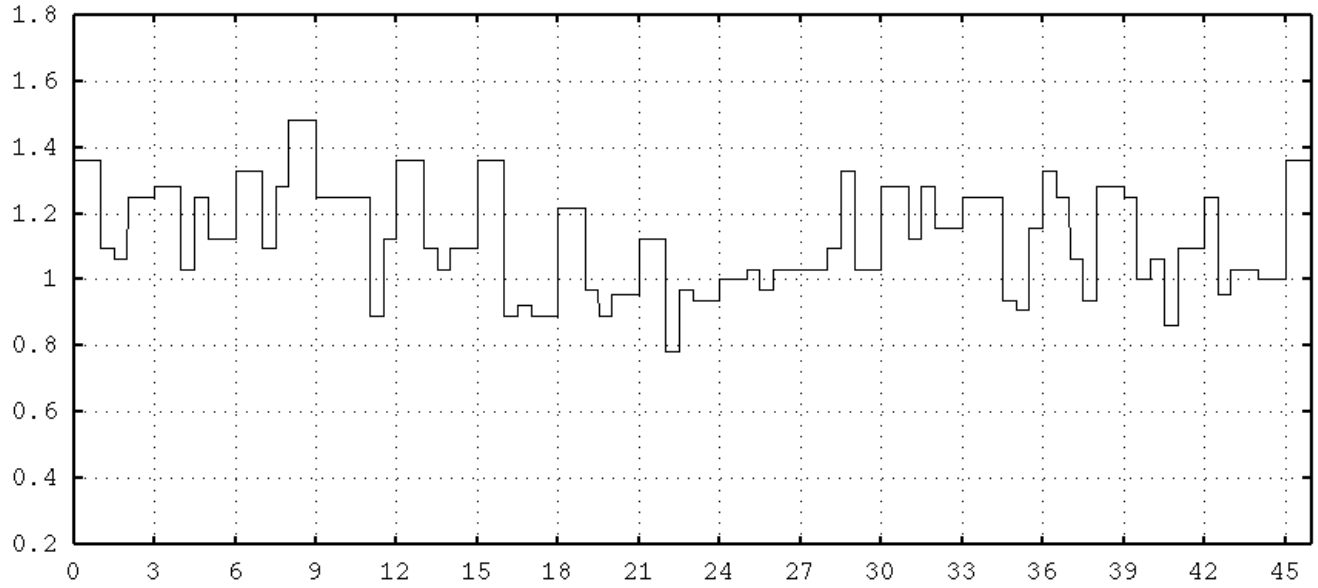

Figure 8: Chopin Waltz op.18, E♭ major, as played by **author** *(dynamics)*.

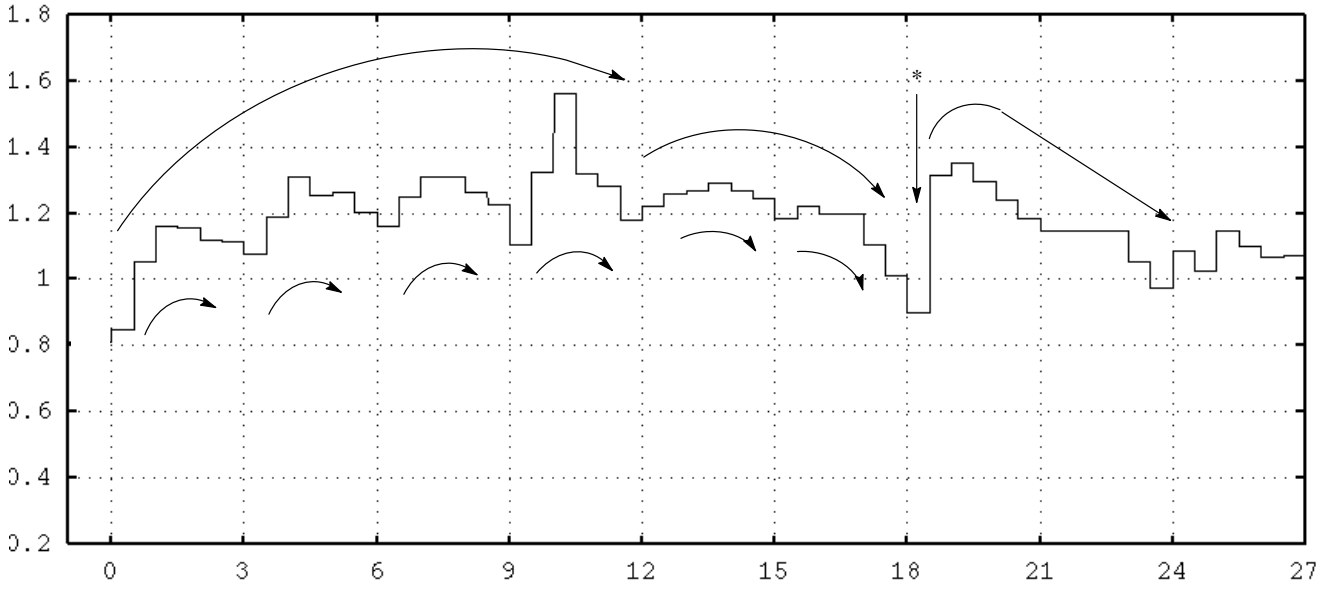

Figure 9: Chopin Waltz op.64, no.2, C♯ minor, as played by **learner** *(tempo)*.

6 Summary and Discussion

This paper has presented research in the intersection of Artificial Intelligence and Art. We have described an implemented system that learns to solve a complex musical task from a surprisingly small set of training 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 with this paper—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 (translating the problem from the *environment representation* to a *learning representation*, in the terminology of Flann and Dietterich, 1986) is an interesting alternative to 'standard' knowledge-based learning, where learning proceeds at the level of the original data, and the knowledge is used to bias induction towards plausible generalizations. 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 an explicit partial *domain theory* that specifies relevant dependencies between various domain parameters. This will require no changes to the system itself, because the learning algorithm 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 (1977) 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 (directed melodic lines, rhythmic gap fills, etc.) are equally important and determine or explain to a large extent the micro-structure 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.

For music as an active art domain, the research points to possible ways of building flexible tools and instruments that can adapt to an artist's performance style. Interactive music programs and instruments are now beginning to be developed and used in performances of electronic music (Rowe, 1993), and systems capable of on-line learning might lead to new artistic possibilities for composers and performers. That will require more research in the direction of incremental, real-time learning. Also, our current repertory of musical structures is limited to classical tonal music and is by no means complete; here, more music-theoretic research will be needed.

References

Aha, D., Kibler D., and Albert, M. (1991). Instance-Based Learning Algorithms. *Machine Learning 6(1)*, pp.37–66.

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 Stone, C. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Clarke, E. (1987). Levels of Structure in the Organisation of Musical Time. *Contemporary Music Review 2(2)*, pp.211–238.

Flann, N. and Dietterich, T. (1986). Selecting Appropriate Representations for Learning from Examples. In *Proceedings of the Fifth National Conference on Artificial Intelligence (AAAI-86)*, Philadelphia, PA.

Lerdahl, F. and Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.

Narmour, E. (1977). *Beyond Schenkerism: The Need for Alternatives in Music Analysis*. Chicago, Ill.: Chicago University Press.

Pazzani, M. and Kibler, D. (1992). The Utility of Knowledge in Inductive Learning. *Machine Learning 9(1)*, pp. 57–94.

Rowe, R. (1993). *Interactive Music Systems: Machine Listening and Composing*. Cambridge, MA: MIT Press.

Sloboda, J. (1985). *The Musical Mind: The Cognitive Psychology of Music*. Oxford: Clarendon Press.

Todd, N. (1989). Towards a Cognitive Theory of Expression: The Performance and Perception of Rubato. *Contemporary Music Review, vol. 4*, pp. 405–416.

Widmer, G. (1993). Plausible Explanations and Instance-Based Learning in Mixed Symbolic/Numeric Domains. In R.S.Michalski and G.Tecuci (eds.), *Proceedings of the Second International Workshop on Multistrategy Learning*, Harper's Ferry, W.VA.