Learning Expression at Multiple Structural Levels

Géaard Widmer
Department of Medical Cybernetics and Artificial Intelligence, University of Vienna, and Austrian Research Institute for Artificial Intelligence, Vienna, Austria
guhard@ai.tuwien.ac.at

Abstract
This paper describes research in the area of Artificial Intelligence / machine learning and tonal music. An implemented computer program is presented that learns expressive performance (dynamics and rubato) from actual recordings by human musicians. The guiding hypothesis is that expression is intimately tied to musical structure as it is perceived by performers and listeners. The most important and novel aspect of our approach is that expression is analyzed and learned at multiple structural levels. The system learns to associate expressive patterns with musical structures at various levels and is then able to apply expressive gestures to new pieces to produce sensible and highly structured expressive performances.

1 Introduction
The work described in this paper continues a series of research projects whose goal is to study the potential of intelligent computer programs to learn expressive performance from examples of human performances. While there are many possible dimensions to musical expression (e.g., rubato, dynamics, articulation, vibrato, various timbral effects, etc.), we concentrate for the moment on dynamics (variations of loudness) and rubato (expressive timing). More precisely, the goal is to develop computer programs that learn general expression rules by "listening" to real performances and can then apply these rules to new pieces to produce musically sensible interpretations.

The motivation for this research is mainly theoretical: we want to study, with the help of Artificial Intelligence methods, what regularities and constraints govern the phenomenon of musical expression, what assumptions about music and musical listening are necessary or helpful for expression to become explainable and thus effectively learnable. In the future, one might also imagine some possible practical applications of the resulting learning methods.

One thing seems clear, and that is in fact the central hypothesis on which this work relies: an important function of expressive performance is the communication of musical structure. The performer uses expressive devices to mark structural aspects he perceives in the music and wants the listener to comprehend. This aspect has been stressed by many music researchers (see, e.g., Clarke, 1988; Sloboda, 1985; Sundberg et al., 1991; Todd, 1989). The conclusion we draw from this is that a computer must possess knowledge about musical structure if it is to be able to comprehend musical expression in any way. In other words, the computer must be capable of "structural hearing." Formulating such explicit models of structural hearing has been a central concern in all our AI & music projects (see, e.g., Widmer, 1992a,b, 1993a, 1994) and also plays a central role in the work to be described here.

The following pages will describe an implemented computer program that learns general rules of musical expression (for the dimensions of dynamics and rubato) from example performances by human musicians. For technical reasons, the input is currently limited to single lines, i.e., melodies without accompaniment. We will first briefly review our last project that dealt with the same task. The problems encountered there motivated a completely new approach, which will be described in section 2.

2 Previous work: learning at the note level
Previous research (Widmer, 1993a; 1994) is also based on the structure-expression hypothesis, had produced a working system that used a qualitative model of structural hearing to learn expression rules at the note level. The structural hearing model, based on concepts from the music theories by Lerdahl & Jackendoff (1980) and Narmour (1977), was intended to provide the computer with roughly the same capabilities of comprehending (the structural aspects of) music as a "normal" human listener. The model, which was really a set of programs, generated analyses of given pieces and provided hints to the learning component to
plausible ‘casual’ connections between the roles a note plays in such a structural analysis and the way we may expect a performer to play it. This information was used by the learning program to distinguish between plausible and implausible hypotheses and to focus on musically sensible generalizations.

The level of granularity of both the learning process and the system’s musical knowledge was the individual note: training examples were the notes in a piece, along with information about how mod. crescendo/diminuendo and accelerando/doloroso/ritardando the performer had applied to them. The musical model annotated each note with the rules it played in the structural analysis (e.g., its metrical strength). The learned rules were also formulated at the note level: for a given note in some new piece, the rules determined what type of variation (crescendo or diminuendo, accelerando or ritardando) should be applied to it, and exactly how much.

Experimental results with relatively simple types of music (e.g., Bach minuets) indicated that the system learned certain expression principles quite effectively, from rather few example performances. In fact, it ‘re-invented’ (through learning from examples) some of the expression rules published years ago by Sundberg and colleagues (1983). Despite this highly interesting result, however, it became clear that the note level is not really appropriate for understanding and learning expression, for a number of reasons:

1) Though the performances produced by the system were in large part musically sensible (at least for the types of music it was tested on), the results lacked a certain smoothness and a sense of both local and global form.

2) It is psychologically implausible that performers think and decide on a purely local level in terms of single notes. It is much more likely that they think in terms of musical chunks (phrases, ideas, and other melodic, harmonic, and rhythmic patterns).

3) Expression is a multi-level phenomenon: if expression is tied to musical structure and form, then expressive shapes, like musical structures, must appear at multiple levels. Local expression patterns may be embedded within larger patterns (e.g., shaping of ornaments within an overall crescendo); patterns may also overlap or conflict. Good performances exhibit structure at both micro and macro levels.

Consequently, we have developed a new approach that abandons the note level and tries to learn expressive principles directly at the level of musical structures. As these structures will be of widely varying scope — some comprising only a few notes, others spanning a number of measures — the system will indeed learn expression at multiple levels at once.

3 Structure-based learning of multi-level expression

The learning scenario is as follows: input to the learning program are the notated scores of melodies along with actual recordings (via MIDI) of these melodies as played by some musician. By comparing the notated piece with the performance data, the system computes dynamics and tempo curves. In the dynamics dimension, the curve is computed as the ratio of the loudness of a note as played vs. the average loudness of the entire piece. To compute the tempo curve, the system tracks the average local tempo, and notes played longer than notated (with respect to the current tempo) are considered instances of ritardando, etc. The desired output are explicit rules that allow the system to decide which variations to apply to new melodies.

The raw training examples as they are presented to the system thus consist of a sequence of notes (the melody of a piece) with numeric values associated with each note that specify the exact dynamics and tempo applied to the note by the performer. However, as observed above, the note level is not adequate. We have thus implemented a transformation strategy which translates the entire learning problem to an abstraction level that is musically plausible and at the same time tractable for the learning program. 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 — various dynamics and rubato patterns — are identified in the example performances. Only such expression patterns are extracted that can be associated with some higher-level musical structures. Learning then proceeds at this abstraction level, and the resulting 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.

3.1 Problem transformation

The problem transformation step proceeds in two stages. First, a musical analysis of the given melody is performed. A set of analysis routines, based on selected parts of the theories by Lerdahl and Jackendoff (1983) and Narcisse (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 1 exemplifies the result of this step with an excerpt from a simple Bach minuet. The musical structures identified here are four measures heard as rhythmic units, three groups heard as melodic units or "phrases" on different levels, two upwardly ascending melodic motifs, 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 expression patterns to serve as examples for the learner are identified. The system tries to find prototypical shapes in the given expression (dynamics and tempo) curves 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: *ascend_level* (no recognizable rising or falling tendency of the curve in the time span covered by the structure), *ascend* (an ascending tendency from the beginning to the end of the time span), *descend*, *ascend降至* (first ascending up to a certain point, then descending), and *descend降至* (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 2 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 1: 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 *descend降至* (decrescendo-crescendo) shape.

### 3.2 Learning

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 musical-theoretic features, of the structure and the notes at its

---

**ICMC Proceedings 1994**

**Expression Analysis**
The desired output of the learning component is a set of general rules that decide, give 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 learning algorithm (Aha et al., 1991) that in effect builds up numeric interpolation tables for each learned symbolic rule to predict precise numeric values. 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.

3.3 Application of learned rules

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. The resulting interpretation can then be played via MIDI on an electronic piano.

4 Experiments

To illustrate the system's performance, this section presents some results from an experiment 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, and Op.70 no.3, played by the author on an electronic piano (Synthesizer) and recorded via MIDI. The results of learning were then tested by having the system play other excerpts from Chopin waltzes.

Figures 3 and 4 show the system's performance of a new piece — beginning of the waltz Op.18 — after learning from the five training pieces. The figures plot the dynamics and tempo curves, 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 score contains some explicit expression marks, 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, of course, compare the system's performance of a piece with a human performance of the same piece and somehow measure the average difference between the two curves. However, the results would be rather meaningless. For one thing, there is no single correct way of playing a piece. Also, relative errors cannot simply be added: some notes 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 sensible. The graphs suggest an 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 two-phrase structure of the piece. In the dynamic dimension, the first and second phrase are played with a recognizable crescendo culminating at the end point of the phrases (the B's at the beginning of the fourth and twelfth measures). 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 literature (e.g., Bodé, 1989).
At lower structural levels, the most obvious phenomenon is the phrasing of the individual measures: 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 the explicit marks in the score (which were not visible to the system). There are clear parallels between the system's dynamics curve and the various crescendo and decrescendo markings and also the p command in measure 5. Two notes were deemed particularly worthy of stress by Chopin (or the editor) and were explicitly annotated with sf; the B♭s at the beginning of measures 4 and 12. Elegantly enough, our program came to the same conclusion and emphasized these 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 figures 3 and 4.

Just to comparison, figure 5 shows the dynamics curve from an independent recording of the same piece by the author. There are strong simil...
Fig. 5: Chopin Waltz op.18, Eb major, as played (before) by the author \textit{(dynamics)}

Fig. 6: Chopin Waltz op.64, no.2, C\# minor, as played by learner \textit{(tempo)}

Expression Analysis

ICMC Proceedings 1994
5 Conclusion

To summarize, this paper has described a computer program that learns general principles of musical expression from example performances by human musicians. Preliminary experiments indicate that the system is capable of learning and applying expressive gestures at different structural levels, thus producing musically sensible performances that exhibit structure both at micro and macro levels. This is achieved by translating the training input and the entire learning task onto an abstraction level that makes relevant musical structures visible, and looking for patterns in the performance curves that can be associated with such musical structures. The basis of this structurally oriented system are two established theories of music (Lerdahl & Jackendoff, 1983; Narmour, 1977). The results can be taken as another confirmation of the hypothesis that expressive patterns are closely linked to musical structure as it is perceived by performers and listeners. Indirectly, the positive results also provide support for the adequacy of the structures postulated by the underlying music theories.

Related work concerning computer models of musical expression has been discussed extensively in (Widmer, 1993a, 1994). Here, we note that technically, the expressive shapes learned by our program are related to Desain and Honig's (1992) time function: a single shape (e.g., a crescendo-decrescendo applied to a musical passage) is a simple time function that controls the amplitude or relative tempo of a passage. The mechanism used for combining such time functions, however, is different (cruder) in our system.

This strategy for combining shapes of different levels is one obviously unsatisfactory aspect of the current model. Simple averaging can lead to blurring effects and generally ignores the fact that there may be a very complex interplay between different levels and also between different expressive dimensions. For example, one might well imagine that an increase in intensity can be expressed as a crescendo, but sometimes also as a decrescendo, depending on the rubato/timing dimension. This is a very complex problem and needs a lot more music-theoretic study.

In addition, the system should be extended to learn also articulation. The distinction between various forms of staccato and legato contributes much to a musically satisfying performance. Chopin Waltzes are a case in point. Whether articulation is best learned at the note level or at higher structural levels remains to be investigated.

Acknowledgments

I would like to thank Prof. Robert Trapp for his continuing support of this research. Financial support for the Austrian Research Institute for Artificial Intelligence is provided by the Austrian Federal Ministry for Science and Research.

References


