Large-scale Induction of Expressive Performance Rules: First Quantitative Results

Gerhard Widmer
Department of Medical Cybernetics and Artificial Intelligence, University of Vienna, and
Austrian Research Institute for Artificial Intelligence, Vienna
gerhard@ai.univie.ac.at

Abstract
The paper presents first experimental results of a research project that aims at identifying basic principles of expressive music performance with the help of machine learning methods. Various learning algorithms were applied to a large collection of real performance data (recordings of 13 Mozart sonatas by a skilled pianist) in order to induce general categorical expression rules for tempo, dynamics, and articulation. Preliminary results show that the algorithms can indeed find some structure in the data. It also turns out that meter and global tempo have a strong influence on expression patterns. Finally, we briefly describe an experiment that demonstrates how machine learning can be used to study some specialized questions.

Introduction
In this paper, we present first quantitative results of a long-term research project that aims at identifying and studying basic principles of expressive music performance with the help of Artificial Intelligence (in particular, inductive machine learning) methods. In contrast to, e.g., Johan Sundberg's Analysis-by-Synthesis approach (Sundberg et al., 1991), our research strategy might be called Analysis by Machine Induction. That is, we develop programs that try to discover general performance principles, by analyzing and learning from real performances by human musicians.

Previous research [e.g., Widmer, 1995, 1996] has shown the general viability of this approach. However, these early studies were extremely limited in terms of data. In fact, they were mostly based on performances by the author himself. Now, for the first time, we managed to collect a truly large corpus of real performance data. This paper reports on first results of machine learning experiments on this data collection. In this initial study, we restrict ourselves to trying to predict the performer's categorical performance choices at a local note-to-note level. That is, we will try to learn rules that predict, for each single note in a piece, whether the performer will apply a crescendo or a decrescendo etc., at that point. Learning expression patterns at higher structural levels is a goal for future research.

The Training Data
The data used in the present study consists of recordings of 13 complete piano sonatas by W.A. Mozart [K.279, 280, 281, 282, 283, 284, 330, 331, 332, 333, 457, 475, and 533], performed by a skilled Viennese concert pianist on a Boesendorfer SE290 computer-monitored concert grand piano. The piano measurements (hammer impact times and pedal movements) were transformed into MIDI format. In order to compute the expressive aspects of the performances, i.e., tempo and dynamics curves and articulation, it was also necessary to code the written score (or at least the notes as notated) in computer-readable form. As entering all this information by hand was out of the question, we decided to try to reconstruct as much as possible of the original score information from the MIDI files. That involves a number of non-trivial problems, such as the identification of elementary and composite musical objects (chords, arpeggios, trills, etc.), beat induction, beat tracking, onset quantization, duration quantization, and pitch spelling, for which we had to develop an ensemble of new heuristic algorithms (Cambourpoulos, 2000; Dixon & Cambourpoulos, 2000). The output of these algorithms was then manually corrected by music students.

The resulting data consists of about 106,000 performed notes, along with information about the ideal (notated) note onsets, durations, metrical information, and annotations (e.g., which notes constitute the melody). From this, the details of timing, dynamics, articulation, and pedaling can be computed.

The experiments to be described in the following sections were performed on the melodies only. Using the full polyphony information will be a topic for future research. For various musical and technical reasons, we eliminated a few pieces (in particular, the "Fantasie" K.475 and all the repeats) from the collection; the resulting training set (melody notes only) comprises 30,332 notes. Each note is described by a number of attributes that represent both intrinsic properties (such as scale degree (an abstraction of pitch), duration, metrical position) and some aspects of the local context (e.g., melodic properties like the size and direction of the intervals between the note and its pre-
deceessor and successor notes, and rhythmic properties like the durations of surrounding notes and some abstractions thereof. We also added global properties of the respective piece (mode, meter, absolute tempo).

An Initial Experiment

The first question to be studied is the following: are there any regularities in the data at all that can be detected and characterized by a machine learning algorithm? How predictable are the performer’s interpretation decisions at a note-to-note level? To simplify matters, we will not look at numeric prediction (i.e., exactly how long or loud is a note played), but rather at categorical decisions. The target classes we wish to predict are defined as follows: in the tempo dimension, a note is assigned to class accelerando if it was played faster than the tempo at the previous note, ritardando if slower. In dynamics, the classes are crescendo if the note was played louder than its predecessor, decrescendo otherwise. And in articulation, we (rather arbitrarily) defined three classes: staccato if a note’s ratio of performed vs. notated duration is less than 0.25, legato if this ratio is greater than 0.95, and portato for all ratios in between.

We initially tested many different learning algorithms that induce different types of models (e.g., nearest neighbor methods, Bayes classifier, decision tables, decision trees, and classification rules). The results generally indicated that decision tree and classification rules perform best (in terms of classification accuracy). An additional advantage of these methods is that they produce intelligible models. Thus, from that point on, we focused on one inductive learning algorithm: the decision tree learner J48, as provided in the WEKA learning toolbox (Witten & Frank, 1999).

The classification accuracies of the decision trees learned by J48 on the Mozart melodies are shown in Table 1. They were measured by 5-fold cross-validation (i.e., training the algorithm on 4/5 of the data, testing the predictive accuracy of the resulting tree on the remaining 1/5, and doing this 5 times in a circular fashion). To assess whether these numbers are good or bad, we compare J48’s prediction performance to the so-called default accuracy, which is the success rate one would achieve by always predicting the most frequent class in the data.

Without going into more details here, we observe that there is indeed something that can be learned: J48 learns to predict the expression categories markedly better than default accuracy, and the differences are statistically significant. We take this as first encouraging evidence for the viability of our research approach.

The next question we investigate is whether there are systematic differences between pieces of different tempo, and pieces with different metrical characteristics (i.e., time signature). If so, we should be able to achieve higher prediction accuracy by training the learning algorithm separately on pieces with different characteristics. And indeed, this turns out to be the case. Table 2 shows the result of an experiment where the training data was first divided into slow and fast pieces, and these were again divided into subgroups according to the time signature of the pieces. Especially this latter finer subdivision leads to a significant increase in the learner’s ability to predict the performer’s categorical decisions. The average accuracy rises from 58% to more than 64% in the tempo dimension, from 62% to 65% in dynamics, and from 67% to over 72% in the articulation domain. Global tempo and meter thus seem to have a strong influence on typical expressive patterns (at least at a note-to-note level).

Identifying Relevant Factors: Automated Feature Selection

Besides global factors like tempo and meter, what are the local characteristics of a musical situation that are most predictive with respect to how a note will be played? One way of approaching this question is to inspect the learned decision trees (or the classification rules derived from them). A more systematic way is to use a set of machine learning methods that perform an automated feature selection. These algorithms attempt to establish which attributes are indeed relevant for class prediction.

In experiments with various feature selection algorithms an interesting distinction emerged: the attributes identified as most relevant for predicting categorical tempo changes were generally related to the

<table>
<thead>
<tr>
<th>tempo</th>
<th>dynamics</th>
<th>articulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>slow</td>
<td>50.14</td>
<td>60.30</td>
</tr>
<tr>
<td>fast</td>
<td>50.73</td>
<td>59.95</td>
</tr>
<tr>
<td>sl 3/4</td>
<td>50.17</td>
<td>64.36</td>
</tr>
<tr>
<td>sl 4/4</td>
<td>50.21</td>
<td>71.55</td>
</tr>
<tr>
<td>sl 3/8</td>
<td>51.75</td>
<td>64.63</td>
</tr>
<tr>
<td>sl 6/8</td>
<td>51.76</td>
<td>68.56</td>
</tr>
<tr>
<td>t. 2/4</td>
<td>50.38</td>
<td>60.60</td>
</tr>
<tr>
<td>t. 3/4</td>
<td>50.40</td>
<td>62.69</td>
</tr>
<tr>
<td>t. 4/4</td>
<td>50.21</td>
<td>62.88</td>
</tr>
<tr>
<td>t. 3/8</td>
<td>50.15</td>
<td>62.82</td>
</tr>
<tr>
<td>t. 6/8</td>
<td>50.62</td>
<td>60.84</td>
</tr>
</tbody>
</table>

Table 1: Classification accuracies (in %) for categorical tempo (accelerando/ritardando) and duration changes (crescendo/decrescendo), and articulation (staccato/portato/legato).

<table>
<thead>
<tr>
<th>tempo</th>
<th>dynamics</th>
<th>articulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>slow</td>
<td>60.30</td>
<td>55.23</td>
</tr>
<tr>
<td>fast</td>
<td>60.60</td>
<td>51.87</td>
</tr>
<tr>
<td>sl 3/4</td>
<td>64.36</td>
<td>52.30</td>
</tr>
<tr>
<td>sl 4/4</td>
<td>71.55</td>
<td>51.87</td>
</tr>
<tr>
<td>sl 3/8</td>
<td>64.63</td>
<td>52.30</td>
</tr>
<tr>
<td>sl 6/8</td>
<td>68.56</td>
<td>52.44</td>
</tr>
</tbody>
</table>

Table 2: Accuracies (in %) for specific types of pieces.
Table 3: Accuracies for different attribute subsets.

<table>
<thead>
<tr>
<th></th>
<th>tempo</th>
<th>dynamics</th>
<th>articulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>rhythm only</td>
<td>56.31</td>
<td>56.71</td>
<td>60.48</td>
</tr>
<tr>
<td>contour only</td>
<td>53.33</td>
<td>62.28</td>
<td>64.06</td>
</tr>
</tbody>
</table>

rhythmic content of the music (specifically, duration ratios between neighboring notes, metrical strength, and meter), while for dynamics and articulation, melodic aspects like local melodic contour and the intervals connecting the note to its neighbors appeared at the top of the list (but meter was also selected).

In order to verify that there is indeed substance to this distinction, we performed the following experiment: two variants of the training dataset were produced, one involving rhythmic, the other involving melodic (i.e., contour-related) attributes only. 448 was then applied to learn tempo, dynamics, and articulation rules from these. The results (in terms of predictive accuracy) are summarized in table 3. They seem to confirm the distinction. Tempo turns out to be more easily predictable from the rhythm-only data, while the dynamics and articulation models benefit more from the melodic contour attributes. Of course, the accuracies are generally a bit lower than the corresponding figures in table 1, where all the attributes were available to the learner — with the exception of the dynamics/contour-only combination, which shows a slight improvement over learning with all attributes; the difference is not statistically significant, however.

This is only a first, preliminary result that needs to be followed up with more systematic research. We will use it as a starting point for future studies that will investigate the role of different structural dimensions in more detail.

A Specialized Experiment: Shorten or Lengthen Short Notes?

While the experiments discussed above do show that regularities can be extracted from our performance data by machine learning algorithms, it is clearly unreasonable to expect that all or even most relevant expressive patterns will emerge at the level of individual notes. For instance, a slowing down that is part of a larger ritardando (for instance, towards the end of a phrase) cannot be explained by reference to a single note and its intrinsic properties. For a machine to be able to learn — and indeed to represent — such patterns, it will have to have a notion of higher-level musical structure; we have not yet succeeded in implementing that.

Our current methods seem better suited to studying those types of expressive phenomena that can be attributed to, and thus described with reference to, local features of the music, such as local stresses on particular notes or specific ways of shaping a particular rhythm. Let us demonstrate the potential of this approach by way of another experiment. As a point of departure, we take the well-known set of performance rules developed over the past 15 years by Johan Sundberg and co-workers (e.g., Sundberg et al., 1991; Friberg, 1991).

In (Friberg, 1991), two rules are proposed that modify note durations and that can be seen to be in direct conflict in certain situations: the Durational Contrast rule generally shortens short notes (notes with duration < 600 ms), by a factor inversely proportional to the note’s duration. The Double Duration rule, on the other hand, lengthens short notes if they are preceded by a note exactly twice as long and followed by a longer note (e.g., a sequence 1/4, 1/8, 3/8).

If we filter the soprano lines of our Mozart sonatas for short notes (notes no longer than the beat level) appearing between two longer ones, we obtain 525 such passages. It turns out that in 232 (42.5%) of these, the ‘sandwich’ short note is actually played longer than the played duration of its predecessor (i.e., the local “tempo” at this point) would suggest; in 57.5% of the cases, the short note is lengthened by the pianist. It is now a straightforward idea to use our learning algorithm to find rules that distinguish between these two types of situations and tell us under what conditions the short note tends to shorten or lengthen short notes, respectively.

Automatic feature selection chooses mainly attributes that relate to metrical aspects and to melodic contour. The time signature of the pieces is also selected as a relevant factor. Reducing the 525 examples to these attributes and applying Weka’s classification rule learner PART (Witten & Frank, 1999) on the data yields a set of 43 classification rules, 17 of which cover a substantial number of examples (at least 10). An analysis of these rules reveals a number of interesting classes of situations. Take, for instance, the following two discovered rules:

- prev_dur_ratio > 2 AND prev_dur_ratio <= 3 AND next_dur_ratio > 0.2 AND meter = 6/8 AND int_next <= 2 ==> accelerating
- prev_dur_ratio <= 2 AND next_dur_ratio > 0.25 AND meter = 6/8 ==> ritardando

The first one describes a class of situations where the pianist shortened the short note relative to the previous one. (prev_dur_ratio is the duration ratio of previous to current note, next_dur_ratio is the duration ratio between current and next note, and int_next is the interval between current and next note, measured in semitones.) Locating the corresponding passages in the sonatas, we notice that all of them are note triples with a duration ratio of 3:1:2 in pieces of 6/8 time as in the theme of the well-known A major sonata K.331.
This is generally known as a *Siciliano rhythm*, and the shortening called for by our learned rule corresponds closely with observations made by Gabrielson [1987].

The second rule, on the other hand, calls for a *lengthening* of the short note and evidently constitutes a special case of Friberg's *Double Duration* rule for 6/8 meters. Incidentally, the distinction expressed by the above two rules seems to confirm a hypothesis by Johan Sundberg [1993], who predicted a lengthening and shortening of short notes in 2:1 and 3:1 duration contexts, respectively, based on considerations of *categorical perception* (i.e., the need to make these two rhythmic patterns more easily distinguishable for listeners) — see also [Parnicutt, 1994].

The strong effect of *meter* on the timing of short notes is demonstrated by a number of other rules, like the following three:

\[
\begin{align*}
\text{meter} &= 3/8 \ AND \\
\text{dur} &= 0.5 \implies \text{accelerando} \\
\text{meter} &= 2/4 \ AND \\
\text{dur} &= 0.125 \implies \text{accelerando} \\
\text{staccato} &= \text{no} \ AND \\
\text{int\_next} &= 1 \ AND \\
\text{meter} &= 4/4 \implies \text{ritardando}
\end{align*}
\]

By the way, the following surprisingly simple rule also emerged:

\[
\text{followed\_by\_trill} = \text{yes} \implies \text{ritardando}
\]

This may point to a general principle of preparing trills by a slight delay.

To summarize, we believe that this type of machine learning experiment can be a useful tool for studying specialized questions relating to local timing or dynamics. The extent to which the discovered rules are style- and/or performer-specific will have to be established in experiments with different types of music (e.g., pieces from the Romantic period) and different performers.

**Conclusion**

Obviously, what has been presented here is only a first step that barely scratches the surface of the problem. Expressive performance is a complex phenomenon, and our performance data supports a rich variety of analytical studies. Our future research will aim at both refining the studies at the note level and extending the investigation to higher structural levels. An immediate goal is to extend the representation language, that is, adding a number of musically relevant concepts to the description of the notes. In particular, *harmony* and *phrase structure* are clearly relevant to expression but have as yet not been used because we lack reliable algorithms for computing them.

In the above experiments, we learned only qualitative models that predict categorical choices like *crescendo* or *decrescendo*. Clearly, such models are of limited value. The final research goal is to obtain *quantitative models* that predict the exact levels of tempo or dynamics. These can then also be used to produce artificial performances.

Substantial improvements in the quality of the models may be expected by moving from the note level to more abstract structural levels (see also Widmer, 1996). One goal will be to identify phenomena at different levels and combine the resulting models. Other issues to be studied in the more distant future include polyphonic aspects, *cross Accentuation*, the chord asynchronicity, and the possibly very intricate interdependencies and interactions between different expression dimensions (e.g., timing and dynamics).

**Acknowledgments**

This research is part of the START program Y99-INF, financed by the Austrian Federal Ministry for Education, Science, and Culture. I would like to thank Roland Batik for allowing us to use his performances, and the Boesendorfer company and in particular Fritz Lachnit for providing the data and technical help. Thanks also to Richard Parnicutt for helpful comments.

**References**


