Machine Discoveries: A Few Simple, Robust Local Expression Principles

Gerhard Widmer
Department of Medical Cybernetics and Artifcial 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 a new approach to discovering general rules of expressive music performance from real performance data via inductive machine learning. A new learning algorithm is briefly presented, and then an experiment with a very large data set (performances of 13 Mozart piano sonatas) is described. The new learning algorithm succeeds in discovering some extremely simple and general principles of musical performance (at the level of individual notes), in the form of categorical prediction rules. These rules turn out to be very robust and general: when tested on performances by a different pianist and even on music of a different style (Chopin), they exhibit a surprisingly high degree of predictive accuracy.

1 Introduction

Expressive performance is a phenomenon of central importance in our musical world. After the pioneering work by Seashore and his collaborators (Seashore, 1936), there has been renewed interest in this phenomenon in the last two decades, with a wealth of empirical work (e.g., the various studies by Gabrielsson (1994, 1999), Gabrielsson & Juslin (1996), Shaffer (1980), Shaffer et al. (1985), Palmer (1988), Repp (1992, 1998, 1999), Sundberg et al. (1991), Sundberg (1993), Friberg (1995), Bresin (2000), Timmers et al. (2000), Canazza et al. (1997), Windsor & Clarke (1997), Windsor et al., 2000, to cite just a few). The goal of all this work is to understand just what it is that performers do to make music 'come alive'.

The predominant approaches to empirically studying expressive performance that have been followed so far are (a) statistical analysis (e.g., Repp, 1992), (b) mathematical modeling (e.g., Todd, 1989, 1992), and (c) *analysis-by-synthesis* (e.g., Friberg, 1991, 1995; Sundberg et al., 1991). In all of these research strategies, it is the human analyst who devises a theory or mathematical model of aspects of expression and then tries to establish the model's empirical validity by testing it on real performance data.

We propose to complement these research strategies with an approach we might call *machine induction*, where a computer uses intelligent data analysis methods (mostly from the area of Artificial Intelligence) to autonomously discover new and potentially interesting regularities and performance principles from expert performance data, i.e., a given set

of example performances. The research areas of *machine learning* (Mitchell, 1997) and *data mining* (Witten & Frank, 1999) have produced a wealth of algorithms that can search for and discover complex dependencies and regularities in extremely large data sets, and can describe their discoveries to the user in intelligible terms. The advantage of such an approach is that the computer is free of any musical preconceptions and expectations and thus might more easily come up with novel and possibly surprising hypotheses.

Also, we believe that it is essential to base such studies on really *substantial* sets of *real-world* performance data. Most studies of performance to date have used rather small sets of carefully selected (subsections of) test pieces. Data Mining methods make it possible to analyze huge amounts of empirical data. In fact, the data sets we are using in our current studies are extremely large, on the order of tens or even hundreds of thousands of performed notes (Widmer, 2001a). We believe that this adds a new quality and empirical validity to the experimental results we are going to present here.

In previous research (Widmer, 2000), we had presented quantitative evidence that it is indeed possible for machine learning programs to find a certain amount of structure in such complex data. In an experimental study with an extended collection of piano performances (13 complete Mozart sonatas), various machine learning algorithms learned to predict the performer's expressive choices (e.g., whether to lengthen or shorten a particular note) at a local, note-to-note level. It was shown that the learners were able to predict the performer's choices with better than chance probability.

However, the learned models were extremely complex. For instance, a decision tree discriminating between *accelerando* (note shortening) and *ritardando* (note lengthening) with 58.09% accuracy had 3,037 leaves (which corresponds to 3,037 classification rules)! That is clearly not desirable as the purpose of our research is to discover intelligible rules that provide new insight.

A detailed analysis showed that this was due to an inappropriate choice of learning algorithm: decision tree learners try to find a *complete model* that tries to explain all of the observed variations on the basis of the given data. Obviously, the level of individual notes is not sufficient as the sole basis for a complete model of performance. Some of the performer's choices may well be explainable by reference to single notes and their local context, but others may only become predictable if one relates them to higher levels of the musical structure or more abstract expression patterns, neither of which were included in the representation of the data.

As a consequence, we have developed a new approach, which is the topic of the current article. We still remain at the note level, but we abandon the idea of complete coverage/discrimination and instead aim at finding *partial models* that 'only explain what can be explained' at the note level. That is, the goal is to identify those aspects of expressive variation that can be reliably characterized and learned at the note level and to learn partial models of these aspects only, which gives small sets of rules that are general and robust, but cover only a fraction of all observed deviations. Such simple, high-coverage rules, while not constituting a complete model of expressive performance, may form the nucleus of robust local expression models which could then be combined with higher-level models of phrasing etc.

The present article describes a comprehensive set of experiments with this new approach. We apply a new rule learning algorithm (specifically designed and implemented for this purpose) to a very large set of expert performance data (13 complete piano sonatas

by W.A. Mozart; some 106.000 performed notes; some 4 hours of music). The goal is to learn rules that describe and predict local *timing*, *dynamics*, and *articulation* changes applied by the pianist. In the experiments it turns out that one can discover small numbers of very simple rules that still cover a substantial number (but by far not all) of the instances of a given category of expressive variation (e.g., note lengthening) and distinguish them quite well from the opposite classes (e.g., shortening). For instance, we will show that 4 simple rules are sufficient to cover 22.89 % of the instances of note lengthening in our large data set.

The validity of the rules thus discovered is assessed empirically in a number of ways. In a first quantitative analysis, we measure the *degree of fit* (in terms of coverage and accuracy) of the rules on the 13 given training performances. That gives an indication of the coverage of the rules, in other words, how much of the expressive variation observed in these particular performances can be 'explained' by these learned rules. The *generality* of the discovered principles is then assessed by measuring their predictive accuracy on (a) performances of Mozart sonatas by a *different pianist*, and (b) performances of pieces of a *different musical style* (Chopin), by a variety of different pianists. Especially the latter analysis reveals that the rules seem to be extremely general and robust.

2 The Training Data

2.1 The Performances

The performance data used for the present study are 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 (Roland Batik) on a Bösendorfer SE290 computer-monitored concert grand piano. The piano measurements (hammer speed and impact times and pedal movements) were transformed into Midi format. In addition to the performance data, we also coded the written score (or at least the notes as notated) in computer-readable form. This was done in a semi-automatic fashion, with the help of a number of heuristic algorithms (Cambouropoulos, 2000). The resulting data consists of some 106,000 performed notes, along with information about the nominal note onsets, durations, metrical information, and annotations (e.g., which notes constitute the melody, which ones are marked *staccato*, etc.). From these two sources — score and performance information — the details of timing, dynamics, and articulation can be computed. To our knowledge, this is by far the most substantial body of expert performance data ever used in empirical performance studies.

The present study restricts itself to an analysis of a single line — the *'melody'*. In most cases, it was clear which notes constitute the melody (mostly the soprano line); sometimes, however, we had to make rather arbitrary choices. The resulting set of melody notes extracted from our data comprises 52,253 notes. From this, we excluded certain types of notes that tend to produce artifacts in the observed performance curves (e.g., grace notes, trills, and arpeggios), which gives an effective training set of 41,116 notes.

When sections were repeated by the pianist (which is the case for 89 out of the 128 sections that make up the sonatas), we included both repeats in the training data, which has a sort of smoothing effect — inconsistent variations (e.g., a lengthening of a note in the first repeat, a shortening of the same note in the second repeat) will tend to be

interpreted as 'noise' by the learner and will not lead to the creation of rules. The effect is similar as if we first averaged the two repeated performances, as done, e.g., in (Repp, 1992).

2.2 The Target Classes

The objective of the present study is to find note-level rules, i.e., rules that predict (or prescribe) how a particular note in a particular context should be played (e.g., longer or shorter than the current tempo would dictate). The expressive dimensions we are currently looking at are (local) timing, dynamics, and articulation. The task of predicting timing etc. is ultimately a numeric prediction problem. One could thus try to apply any of the large number of existing *regression methods* to derive numeric expression models from the data. However, as our goal is to find readily understandable and interpretable principles, we will, in a first step, search for rules that predict *categorical decisions* (e.g., play louder or softer) rather than precise amounts (e.g., the precise level of loudness to be applied to a note). The result will be rules that are immediately interpretable and can be compared to hypotheses formulated by other researchers, most notably Friberg & Sundberg's set of performances rules (e.g., Friberg, 1991).

The note performance categories we wish to predict are defined as follows:

- in the timing dimension, a note N is considered an example of class *lengthen* if the inter-onset interval (IOI) between the start of N and the next note is lengthened relative to (a) the instantaneous tempo at the note's predecessor and (b) the current local tempo (computed over the last 20 events); class *shorten* is defined analogously;
- in dynamics, a note N is considered an example of class *louder* if it was played louder (i.e, with higher MIDI velocity) than its predecessor, and also louder than the average level of the piece; class *softer* is defined analogously;
- in articulation, three classes were defined: *staccato* if a note's ratio of performed vs. notated duration (PDR – *Played Duration Ratio*) is less than 0.8, *legato* if the ratio is greater than 1.0, and *portato* otherwise; we will only try to learn rules for the classes staccato and legato.

A performed note is considered a counter-example to a given class if it belongs to one of the competing classes. Note that by the above definitions, there will be notes that are neither examples nor counter-examples of some category. For instance, an IOI needs to be shortened relative to both the current local (windowed) tempo and the IOI's immediate predecessor for the corresponding note to count as a significant lengthening. This is to reduce the chances of erroneously counting unintentional or random fluctuations as examples of real expressive choices.

The total numbers of training examples (notes pertaining to each category) that result from this are summarized in Table 1, separately for different global tempi[1] and time signatures of the corresponding sonata sections.

[1]The distinction between fast and slow pieces was made simply on the basis of the qualitative tempo indication given in the score — *andante*, *adagio* etc. are classified as *slow*, *allegretto*, *allegro*, and *presto* as *fast*.

	timing		dynamics		articulation		
	longer	shorter	louder	softer	staccato	legato	other
slow 2/2	790	786	677	580	1,374	473	625
slow 3/4	1,209	1,169	1,081	942	1,371	1,326	1,177
slow 4/4	903	916	836	708	961	1,212	821
slow 3/8	157	153	151	100	150	216	125
slow 6/8	598	667	523	469	1,037	458	530
fast 2/2	1,945	1,925	1,662	1,447	3,227	1,305	1,394
fast 2/4	2,148	2,206	1,851	1,376	4,270	887	1,345
fast 3/4	2,048	1,972	1,838	1,467	3,611	1,387	1,289
fast 4/4	2,107	2,078	1,781	1,413	3,486	1,338	1,537
fast 3/8	738	711	623	435	1,387	337	343
fast 6/8	767	724	606	492	1,258	317	542
slow	3,657	3,691	3,268	2,799	4,893	3,685	3,278
fast	9,753	9,616	8,361	6,630	17,239	5,571	6,450
total	13,410	13,307	11,629	9,429	22,132	9,256	9,728

Table 1: Numbers of training examples of various categories.

Regarding articulation, it must be noted that we do not take into account the pedalling. Thus, in particular, the number of *staccato* examples listed in Table 1 is definitely too high; the number of notes 'really' meant to be (and really sounding) *staccato* is probably significantly lower. Also, the thresholds of 0.8 and 1.0 that define *staccato* and *legato* were chosen rather arbitrarily. Our *staccato* cases will thus be a mixture of 'real' *staccati* and what might more appropriately be considered insertions of micro-pauses after certain notes. Further studies regarding the *staccato* issue are under way (Bresin & Widmer, 2000).

Each note is labelled with its corresponding class and described by a number of attributes that represent both intrinsic properties (such as scale degree, 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 predecessor and successor notes, and rhythmic properties like the durations of surrounding notes and some abstractions thereof). The rules presented in the following sections give a flavour of the types of descriptors used.

3 Learning Partial Rule-based Models

From the perspective of machine learning and data mining, the problem is to find partial descriptive models of categories such as 'situations where a note is lengthened' vs. 'situations where a note is shortened'. 'Descriptive' means that the models should characterize classes of situations that are treated in a similar way by the performer, and these descriptions should preferably be simple and musically meaningful; we will use rule-based models and algorithms for learning classification rules from data.

'Partial' means that we do not expect the models to be able to cover and describe all of (or even a large part of) the instances of a given category observed in the data. We are looking for rule models that capture only a (possibly small) part of all observations, but

describe these in meaningful terms and separate them from instances of other categories reasonably well.

The search for partial models necessitates a special approach to rule learning, since this is not the standard type of problem addressed in 'classical' classification settings. Consequently, we have developed a new machine learning algorithm named PLCG that is geared towards finding simple, robust classification rules in complex, noisy data. PLCG is a generic learning algorithm that can be applied to arbitrary domains. The algorithm and its properties are described and analyzed in more detail in (Widmer, 2001b). For the purposes of the present article, it seems sufficient to give a slightly simplified music-specific specification of the procedure as it is actually used in our musical experiments:

For each expression dimension (timing, dynamics, articulation), PLCG does the following:

1. Separate the data into subsets according to tempo (slow/ fast) and time signatures (2/2, 2/4, 3/4, 4/4, 3/8, 6/8) of the pieces.

2. Learn partial rule models from each of these 2×6 subsets separately. This is done with a rule learning algorithm of the 'sequential set covering' variety (Fürnkranz, 1999) that was specially devised and implemented for this project. The resulting sets of rules will most likely be overly specific (specialized with respect to tempo and time signature).

3. Merge all these learned rule sets into one large set.

4. Perform a hierarchical *clustering* of the rules into a tree of clusters of similar rules, according to a syntactic/semantic rule similarity measure specially defined for our application.

5. For each of these clusters, compute the *least general generalization* of all the rules in the cluster (i.e., a generalization that subsumes all the rules and is no more general than necessary). The result is one rule per cluster. The resulting tree represents generalizations of various degrees of the original rules.

6. From this generalization tree, select those rules that optimize a given trade-off function between coverage (the number of cases covered by the rule) and precision (the percentage of covered cases that actually do belong to the rule's predicted class – in other words, the percentage of correct predictions made by the rule).

The goal of the entire procedure is to arrive at rules that both cover a significant number of cases (and thus describe a significant 'sub-concept') and are still reasonably accurate in distinguishing positive examples from counter-examples. This is achieved via the process of learning many specialized rule sets (in step 2 above), finding rules in these sets that seem to describe similar sub-concepts (step 4), generalizing these to varying degrees (step 5), and selecting, from these alternative generalizations, those rules for the final model that optimize some criteria concerning coverage and accuracy. Again, more detail can be found in (Widmer, 2001b).

4 Experimental Results: Some Simple and General Performance Rules

The above procedure was carried out on the complete Mozart performance data set (41,116 notes), for each of the three expression dimensions timing, dynamics, and articulation. The final sets of rules selected (from a total of 383 specialized rules) consist of only 17 rules: six rules for local timing, six for local dynamics, and five for articulation. These seem to represent very general (and mostly very simple) principles, some of which cover or seem to 'explain' a surprisingly large number of the pianist's expressive choices. The following sections discuss the rules in detail and present a quantitative evaluation of their generality and precision. A summary of all the discovered rules is given in Table 2.

4.1 The Discovered Rules

In the following, the rules are listed in the representation language used by the learning system and are briefly paraphrased. The numbers TP/FP following the paraphrase denote the number and percentage of positive examples in the training data that are correctly covered (*true positives TP*) and the number of cases where a rule makes an erroneous prediction, e.g., predicts a lengthening when the pianist actually shortened the note (*false positives FP*); the percentages are relative to the true numbers of positive examples and counter-examples, respectively. π is the rule's *precision* — the ratio of cases (out of the total number of cases in which the rule did make a prediction) where the rule's prediction was correct; in other words, $\pi = TP/(TP + FP)$. We give a separate evaluation of the rules on slow and fast pieces, as there are some substantial differences sometimes.

Timing: lengthening notes (IOIs)

In the domain of local timing, there emerged two rules that appear to represent very strong, general principles, along with a third rule that describes an interesting class of situations. The most general discovered rule is

RULE TL1:
```
lengthen IF
  abstr_dur_context = equal-longer
```

"Lengthen the middle note in a "cumulative" (Narmour, 1977) 3-note rhythm situation (i.e., given two notes of equal duration followed by a longer note, lengthen the note that precedes the final, longer one)."

slow: 1,020 (27.89 %) / 153 (2.59 %), $\pi = .870$
fast: 1,645 (16.87 %) / 752 (5.14 %), $\pi = .686$
all: 2,665 (19.87 %) / 905 (4.40 %), $\pi = .746$

This is an extremely simple principle that is also surprisingly general and quite precise, especially in the slow pieces: there, TL1 covers 1,020 cases (27.89 % of all examples of lengthening in the data) correctly, while predicting a non-observed lengthening in only

153 instances (2.59 % of the cases where the pianist did the opposite). For fast pieces, TL1 predicts a substantial number of positive instances correctly, but makes a higher number of wrong predictions (5.14 %).

A second very simple rule that emerged very strongly and that is obviously related to TL1 is

RULE TL2:
`lengthen IF`
`next_dur_ratio` ≤ 0.334

"Lengthen a note if it is followed by a substantially longer note (i.e., the ratio between its duration and duration of the next note is $\leq$ 1:3)."

slow: 725 (19.82 %) / 132 (2.23 %), $\pi = .846$
fast: 1,063 (10.90 %) / 439 (3.00 %), $\pi = .708$
all: 1,788 (13.33 %) / 571 (2.78 %), $\pi = .758$

A variant of this rule, with substantially higher positive coverage but also more incorrectly predicted negative instances (particularly in the fast pieces), is

RULE TL2a:
`lengthen IF`
`next_dur_ratio` < 1 `&`
`metr_strength` ≤ 2

"Lengthen a note if it is followed by a longer note and if it is in a metrically weak position."

slow: 1,121 (30.65 %) / 246 (4.16 %), $\pi = .820$
fast: 1,651 (16.93 %) / 918 (6.27 %), $\pi = .643$
all: 2,772 (20.67 %) / 1,164 (5.66 %), $\pi = .704$

(In the quantitative experiments to be described below, we will use TL2a for slow pieces only.)

Clearly, rules TL1 and TL2/TL2a are strongly related, but they also partly complement each other; taken together, they cover 2,965 (22.11 %) of the positive examples, which is substantially more than either of them covers in isolation.

It is remarkable that essentially *one* simple principle as embodied in rules TL1 and TL2 — lengthen a note if it is followed by a longer one — is sufficient to account for more than one fifth of all the significant note lengthenings observed in a large body of performance data. We consider this a surprising discovery that merits some more detailed investigations.

Of course, variants of this principle have been observed before. For instance, rules TL1 and TL2 cover all the cases of lengthening the last in a sequence of short notes before a (half) cadence that were observed by Palmer (1996, p.442) in a Mozart performance, and which were interpreted there as a strategy to draw attention to the upcoming cadence.[2]

[2]In that case, instead of *lengthening*, it would probably be more appropriate to speak of *delaying* the following note. The physical effect is the same, of course, but the musical intention (and, indeed, the perception by the listener) may be different.

The particular timing pattern was discussed by Palmer under the heading "Performer-specific Expression". Our discovery shows that at least our pianist applies the same principle, and very consistently so.

The reader should be aware that the specific formulation of a rule found by our learning algorithm is usually but one of many possible ways of characterizing a set of musical situations. It may be that there are alternative conditions that would describe (more or less) the same set of examples. Also, our learner's representation language is (necessarily) limited in many ways. For instance, it currently has no notion of grouping and phrase structure, because these are not captured in our representation of the scores. The learner is thus sometimes forced to "paraphrase", i.e., to use available descriptors to describe classes of situations that would really best be characterized with respect to, e.g., aspects of the phrase structure.

So, it may well be that rules TL1 and TL2 coincide or at least overlap with some local lengthening or delaying principles formulated by other researchers in other terms. For instance, it could be the case that the type of 'cumulative rhythm' referred to by rule TL1 occurs mostly at the end of a phrase or a low-level group (with the long note being the final note), in which case TL1 could just as well be interpreted as describing the preparation of a group ending by lengthening the penultimate note.

Generally, performance researchers may have specific preferences, based on musical experience and music-theoretic knowledge, to explain their empirical findings in terms of particular aspects of musical structure. An inductive learning system has no such music-theoretic 'bias'. The only criteria it seeks to optimize are generality, accuracy, and simplicity of the prediction rules. So one must be careful not to interpret such machine-induced rules too literally. Even if a rule fits the data well, that does not mean that the performer conceptualizes his/her actions or decisions in these terms. It simply means that this is one of the simplest ways of characterizing a class of situations where the performer exhibits a consistent and predictable behaviour.

In any case, the large number of cases seemingly 'explained' by rules TL1 and TL2 is an interesting discovery that deserves further investigation. In addition, there was one weaker, less predictive rule that emerged from the learning process:

RULE TL3:
`lengthen IF`
`dir_next = up &`
`int_next` $>$ `p4 &`
`metr_strength` $\leq$ `2 &`
`int_prev` $\leq$ `maj2`

"Lengthen a note if it precedes an upward melodic leap of more than a perfect fourth, if it is in a metrically weak position, and if it is preceded by (at most) stepwise motion (`int_prev` $\leq$ `maj2`*)."*

slow: 95 (2.60 %) / 38 (0.64 %), $\pi = .714$
fast: 164 (1.68 %) / 94 (0.64 %), $\pi = .636$
all: 259 (1.93 %) / 132 (0.64 %), $\pi = .662$

TL3 obviously represents a tendency rather than a strong rule. It is not as clear-cut as TL1 and TL2 and makes a rather large number of wrong predictions, but it still

distinguishes significantly between cases of lengthening and shortening. TL3 appears noteworthy because it has an interesting parallel in the *articulation* rules described below (staccato rule AS3).

What TL3 describes is a tendency to slightly delay the target note of an upward leap, by lengthening the IOI occupied by the start note of the leap. According to the articulation rule AS3 below, this seems to usually go along with a slight *staccato* (more appropriately: the insertion of a micropause before the target note of the leap), which amplifies the sense of delay and separation.

The preparation of leaps via timing and articulation has also been studied previously. For instance, the KTH rule set (Friberg, 1995) features a pair of rules named *Leap Tone Duration (LTD)* and *Leap Articulation (LA)* that pertain to this type of situation. *Leap Articulation* inserts a micropause between the notes of a leap, and *Leap Tone Duration* shortens the first and lengthens the second note (IOI) of a leap for upward leaps, and does the opposite for downward leaps. Note that rule TL3 learned by our system predicts the *opposite* of Friberg's *LTD* rule — it calls for a *lengthening* of the first IOI in upward leaps (and this is supported by our performance data). This suggests that — at least in Mozart piano music — the *Leap Tone Duration* rule should be used with a negative k parameter. Regarding downward jumps, there does not seem to be a general trend in our set of performances.

Timing: shortening notes (IOIs)

Note shortening and local speedups seem much more difficult to predict. That may seem disappointing, but maybe it is not all that surprising. Introspection tells us that IOI shortenings and local speedups are not something one uses as consciously as an expressive device as the lengthening or delaying of notes or IOIs or a slowing down.

The learning algorithm did not find any really general rule that covers a substantial number of positive instances and is strongly discriminative. Rules with a reasonably large coverage also tend to be overly general, predicting a shortening in many cases where the pianist did not apply one. At best, these rules seem to represent general tendencies that would need to be made more specific in order to be useful as prescriptive rules. One rule with a reasonably positive TP/FP ratio (at least for slow pieces) is

RULE TS1:
`shorten IF`
`  prev_dur_ratio` $\leq$ `0.67 &`
`  next_dur_ratio` $>$ `1.0`

"Shorten a note (IOI) N in a sequence PN-N-NN if it is longer than its predecessor and longer than its successor (more precisely, if the duration ratio PN:N $\leq$ *2:3 and N:NN* $>$ *1:1)."*

slow: 354 (9.59 %) / 175 (2.94 %), $\pi = .669$
fast: 489 (5.09 %) / 527 (3.61 %), $\pi = .481$
all: 843 (6.34 %) / 702 (3.42 %), $\pi = .546$

which expresses a tendency to shorten long notes between shorter ones — a sort of smoothing action. We will use this rule only for slow pieces in the following experiments.

A typical piece where this rule applies is the first section of the sonata K.331, where TS1 together with the lengthening rule TL2a quite accurately reproduces the pianist's timing (see Figure 1 below).

Apart from such weak tendencies, the system was able to discover only a few rather specialized rules. The one with the clearest musical interpretation is

RULE TS2:
shorten IF
  tempo = fast &
  meter = 3/8 &
  prev_dur_ratio $>$ 2.0 &
  dur $\leq$ 0.5 &
  next_dur_ratio $\leq$ 0.99

"Shorten a note (IOI) in fast pieces in 3/8 time if the duration ratio between previous note and current note is larger than 2:1, the current note is at most a sixteenth (its duration is $\leq$ 0.5 beat units), and it is again followed by a longer note."

slow: 0 (0.00 %) / 0 (0.00 %)
fast: 43 (0.45 %) / 4 (0.03 %), $\pi = .915$
all: 43 (0.32 %) / 4 (0.02 %), $\pi = .915$

This rule describes a well-known phenomenon that has been observed in empirical studies before, namely the common shortening of sixteenth notes following a dotted eighth (and usually again followed by a longer note), as in the famous theme of the A major sonata K.331 (Gabrielsson, 1987). Most of the cases covered by rule TS2 do indeed fall into this dotted-eigth–sixteenth note category. Note that in our case, the rule explicitly limits this prediction to fast pieces (as, e.g., in the third movement of the Sonata K.280). In the slow ones (e.g., the beginning of K.331) our pianist tended not to apply this shortening in a consistent manner (in only 16 out of 50 cases did the system find a significant shortening).

Note also that rule TS2 calls for the duration ratio to be *larger than* 2:1. It has been observed by others that duration ratios of 2:1 are usually blurred by performers by *lengthening* the shorter (or shortening the longer) note (see also rule TS1 above). The difference in performing 2:1 vs. 3:1 duration ratios has been linked to considerations of categorical perception (i.e., the need to make these two rhythmic patterns more easily distiguishable for listeners) by some researchers (e.g., Sundberg, 1993; but see also Parncutt, 1994). It is nice to see this principle emerge automatically from actual performances via machine learning.

Dynamics: stressing notes

In the dynamics domain, a few quite clear rules emerged, but with rather low coverage. Some local stresses are very well predictable, but they constitute only a relatively small fraction (11.33%) of the total number of observed stresses. What is noteworthy about the following rules is that they all relate a dynamic stress to a melodic contour peak or at least to upward melodic movement. Again, this kind of link has been made by many researchers (e.g., Palmer, 1996).

RULE DL1:

```
louder IF
  dir_prev = up &
  int_prev > p4 &
  metr_strength > 2
```

"Stress a note by playing it louder if it is preceded by an upward melodic leap larger than a perfect fourth."

slow: 200 (6.12 %) / 36 (0.68 %), $\pi = .847$
fast: 547 (6.54 %) / 172 (1.33 %), $\pi = .761$
all: 747 (6.42 %) / 208 (1.14 %), $\pi = .782$

RULE DL2:

```
louder IF
  mel_contour = up_down &
  int_prev > min3 &
  metr_strength > 2
```

"Stress a note by playing it louder if it forms the apex of an up-down melodic contour and is preceded by an (upward) leap larger than a minor third."

slow: 223 (6.82 %) / 81 (1.52 %), $\pi = .734$
fast: 667 (7.98 %) / 246 (1.90 %), $\pi = .731$
all: 890 (7.65 %) / 327 (1.79 %), $\pi = .731$

RULE DL3:

```
louder IF
  prev_dur_ratio ≤ 0.5 &
  dir_prev = up &
  metr_strength > 3
```

"Stress a note by playing it louder if it is at least twice as long as its predecessor (the duration ratio previous vs. current note is ≤ 1:2), is reached by upward motion, and is in a quite strong metrical position."

slow: 123 (3.76 %) / 120 (2.26 %), $\pi = .506$
fast: 359 (4.29 %) / 147 (1.14 %), $\pi = .709$
all: 482 (4.15 %) / 267 (1.46 %), $\pi = .644$

This last rule, which overlaps somewhat with rule DL1, will be used only in fast pieces in the following experiments.

Dynamics: attenuating notes

Local *diminuendi* seem to be difficult to predict. Three rules emerged, of which one (DS1) exhibits reasonably high coverage (4%), while the other two are more specialized, but also more precise. The first rule pertains to a simple scenario:

RULE DS1:
```
softer IF
  prev_dur_ratio > 5.0
```

"Attenuate a note by playing it softer if it is less than 1/5 the duration of its predecessor."

slow: 159 (5.68 %) / 49 (0.85 %), $\pi = .764$
fast: 218 (3.29 %) / 105 (0.73 %), $\pi = .675$
all: 377 (4.00 %) / 154 (0.77 %), $\pi = .710$

This represents a reasonably clear tendency ($\pi = .710$). Softening short notes after long ones seems to be a reasonable thing to do in many situations; it might be interesting to search for variants of this principle where the duration ratio is smaller than 5:1.

The other two rules that were found predict a softening of notes that terminate certain types of downward leaps:

RULE DS2:
```
softer IF
  dir_prev = down &
  int_prev > maj3 &
  metr_strength ≤ 1 &
  dur_prev > 0.33
```

"Attenuate a note by playing it softer if it is preceded by a downward leap larger than a major third, is metrically weak, and is preceded by a note at least 1/3 of a beat long."

slow: 70 (2.50 %) / 24 (0.42 %), $\pi = .745$
fast: 103 (1.55 %) / 24 (0.17 %), $\pi = .811$
all: 173 (1.83 %) / 48 (0.24 %), $\pi = .783$

RULE DS3:
```
softer IF
  dir_prev = down &
  int_prev > p5 &
  metr_strength ≤ 1
```

"Attenuate a note by playing it softer if it is preceded by a downward leap larger than a perfect fifth and is metrically weak."

slow: 63 (2.25 %) / 12 (0.21 %), $\pi = .840$
fast: 106 (1.60 %) / 27 (0.19 %), $\pi = .797$
all: 169 (1.79 %) / 39 (0.19 %), $\pi = .813$

Rules DS2 and DS3 are quite specialized, linking dynamic attenuation to metrically weak notes reached by downward leaps. Note the parallel to the dynamic stress rules DL1–DL3, which relate a local increase in intensity to upward melodic motion. Again, this seems to make intuitive sense.

Articulation - staccato:

Staccato, defined as any note played with a *Played Duration Ratio (PDR)* ≤ 0.8, turns out to be the most easily predictable dimension. Four quite strong rules emerged. The first of these is rather trivial:

RULE AS1:
`staccato IF`
`  marked_staccato = yes`

"Play a note staccato *if the note is marked with a staccato dot in the score."*

slow: 913 (18.66 %) / 84 (1.21 %), $\overline{x} = 0.36$, $\sigma = 0.16$, $\pi = .916$
fast: 2,158 (12.52 %) / 143 (1.19 %), $\overline{x} = 0.30$, $\sigma = 0.16$, $\pi = .938$
all: 3,071 (13.88 %) / 227 (1.20 %), $\overline{x} = 0.31$, $\sigma = 0.16$, $\pi = .934$

$\overline{x}$ here denotes the mean of the PDR values of the performed notes covered by the rule (i.e., the average degree of staccato); σ is the standard deviation. The only interesting observation about this rule is that the pianist observed only 93.12% of the *staccato* indications pertaining to melody notes in the score;[3] the higher percentage of non-observed *staccati* occurs in the slow sections (8.43% vs. 6.21%).

The next rule is equally simple, but even more accurate:

RULE AS2:
`staccato IF`
`  int_next = unison`

"Play a note staccato *if it is followed by a note of the same pitch (i.e., the interval between the note and its successor is a unison)."*

slow: 1,057 (21.60 %) / 20 (0.29 %), $\overline{x} = 0.35$, $\sigma = 0.14$, $\pi = .981$
fast: 1,872 (10.86 %) / 7 (0.06 %), $\overline{x} = 0.27$, $\sigma = 0.12$, $\pi = .996$
all: 2,929 (13.23 %) / 27 (0.14 %), $\overline{x} = 0.30$, $\sigma = 0.13$, $\pi = .991$

This is again an obvious principle that has been observed by others. In the KTH rule set it appears as a rule named *Repetition Articulation* (Friberg, 1995). Instead of *staccato*, it might be more appropriate to speak of *temporal separation* of notes. The surprising fact is the high precision of this rule ($\pi = .991$): only in 27 out of the total 2,956 cases of tone repetition did the pianist *not* insert a micropause of at least 20 % of the repeated note's duration. This may have physical reasons, being caused by the mechanics of the performer's finger/hand movements and the inertia of the piano action — in fast pieces, only 7 repeated notes (out of 1,879) were played with a PDR > 0.8. In any case, this type of rule will probably have to be a part of any rule system that is to produce musically sounding performances (or at least performances that sound familiar to us).

[3]The score used by the pianist was the Urtext Edition published by G.Henle Verlag, E.Herttrich, ed., Munich, 1977.

The two other rules that emerged as being quite discriminative cover a smaller number of positive examples, but point to two interesting functions of *staccato* (or rather, of inserting micropauses between notes):

RULE AS3:

```
staccato IF
  int_next > p4 &
  dir_next = up &
  metr_strength ≤ 2
```

"Insert a micropause after a note if it precedes an upward leap larger than a perfect fourth and is metrically weak."

slow: 307 (6.27 %) / 161 (2.31 %), $\overline{x} = 0.45$, $\sigma = 0.21$, $\pi = .656$
fast: 930 (5.39 %) / 239 (1.99 %), $\overline{x} = 0.52$, $\sigma = 0.18$, $\pi = .796$
all: 1,237 (5.59 %) / 400 (2.11 %), $\overline{x} = 0.50$, $\sigma = 0.19$, $\pi = .756$

RULE AS4:

```
staccato IF
  next_dur_ratio ≤ 0.4 &
  dir_prev = down
```

"Insert a micropause after a note if it is reached by downward motion and is followed by a note more than twice as long (i.e., the ratio between its duration and duration of the next note is ≤ 0.4*)."*

slow: 375 (7.66 %) / 282 (4.05 %), $\overline{x} = 0.43$, $\sigma = 0.19$, $\pi = .571$
fast: 840 (4.87 %) / 198 (1.65 %), $\overline{x} = 0.48$, $\sigma = 0.17$, $\pi = .809$
all: 1,215 (5.49 %) / 480 (2.53 %), $\overline{x} = 0.47$, $\sigma = 0.18$, $\pi = .717$

It is noteworthy that both AS3 and AS4 have clear parallels in the rules discovered in the timing domain. The insertion of a micropause before an upward leap (rule AS3) seems to usually go along with a lengthening of the corresponding IOI (cf. rule TL3 above). Likewise, rule AS4 has counterparts in the timing rules TL1 and TL2: long notes closing a "cumulative rhythm" (Narmour, 1977) (and often occurring at the end of phrases or lower-level groups) are often emphasized by both delaying them and separating them by a micropause. Both of these tendencies seem to be stronger in fast than in slow pieces (see the respective π values). Also observe that the mean PDR value is significantly higher in these "micropause rules" AS3 and AS4 than in the "genuine" *staccato* rules ($\overline{x}_{AS3} = 0.50$, $\overline{x}_{AS4} = 0.47$ vs. $\overline{x}_{AS1} = 0.31$, $\overline{x}_{AS2} = 0.30$). That also seems to support our distinction between genuine staccato and micropause insertion.

Taken together, the four *staccato* rules presented here account for 6,916 (31.25 %) of all the 22,132 *staccati* observed in our performance data, while making false predictions in only 1,089 (5.73 %) of the cases, which gives a precision of $\pi = .864$; by restricting the applicability of rules AS3 and AS4 to fast pieces only, we can improve this ratio to 6,456 (29.17 %) / 667 (3.51 %) ($\pi = .906$). It seems quite remarkable that such a large proportion of the observed *staccati* can be explained by only four very simple rules.

Articulation - legato:

Legato, defined as an effective overlap between notes (i.e., a *Played Duration Ratio (PDR)* > 1.0) seems the most difficult category to predict. The system could find only one weak 'rule' with a precision > 0.5 for both slow and fast pieces:

RULE AL1:
legato IF
  staccato = no &
  mel_contour = up_down &
  dur $\leq$ 0.334 &
  metr_strength $>$ 2

"Play a note legato *if it is not marked staccato in the score, if it forms the apex of an up-down melodic contour, if it is quite short ($\leq$ 1/3 of a beat), and is metrically quite strong."*

slow: 227 (6.16 %) / 156 (1.91 %), $\overline{x} = 1.35$, $\sigma = 0.55$, $\pi = .593$
fast: 460 (8.26 %) / 436 (1.84 %), $\overline{x} = 1.24$, $\sigma = 0.26$, $\pi = .513$
all: 687 (7.42 %) / 592 (1.86 %), $\overline{x} = 1.27$, $\sigma = 0.39$, $\pi = .537$

This 'rule' should not be taken too seriously. If anything, it points to one potential factor that might be related to legato, namely, the occurrence of a melodic peak — a detailed analysis of the learning process reveals that many of the more specialized rules (with a higher precision) learned in the individual learning runs contained a reference to an up-down melodic contour.

There are at least two reasons why legato is hard to predict. First, the number of instances of *legato* in the training data is much lower than the number of *staccati*. That makes it more difficult to achieve a good TP/FP ratio — any rule covering a substantial part of the *legato* cases is also very likely to erroneously cover a relatively large number of counterexamples (see the TP/FP ratios and corresponding percentages in AL1 above). Second, while *staccato* markings in the score were included in the description of the training data (i.e, the learner knew which notes are marked *staccato*), no comparable score information was coded for *legato*. In particular, the slurs indicating intended phrasing in the score are currently not included in our symbolic, machine-readable representation of the score (for pragmatic reasons to do with the score data entry process). Thus, the learner was totally blind to this potentially important source of articulation information.

4.2 Summary of the Discovered Rules

The complete set of rules is listed once more in Table 2, along with their individual coverage.

5 Quantitative Validation of the Discovered Principles

In this section, the validity of the discovered rules is assessed empirically in three ways. In a first quantitative analysis, we look at the collective *degree of fit* (in terms of coverage

Rule	Action	Conditions	pos. coverage (slow+fast)		Precision slow	fast	total
TL1	lengthen IF	abstr_dur_context = equal-longer	2,665	(19.87 %)	.870	.686	.746
TL2	lengthen IF	next_dur_ratio $\leq$ 0.334	1,788	(13.33 %)	.846	.708	.758
TL2a*	lengthen IF	next_dur_ratio $\leq$ 0.99 & metr_strength $\leq$ 2	1,121	(8.36 %)	.820	—	.820
TL3	lengthen IF	dir_next = up & int_next > p4 & metr_strength $\leq$ 2 & int_prev $\leq$ maj2	259	(1.93 %)	.714	.636	.662
TS1*	shorten IF	prev_dur_ratio $\leq$ 0.67 & next_dur_ratio > 1.0	354	(2.66 %)	.669	—	.669
TS2**	shorten IF	tempo = fast & meter = 3/8 & prev_dur_ratio > 2.0 & dur $\leq$ 0.5 & next_dur_ratio $\leq$ 0.99	43	(0.32 %)	—	.915	.915
DL1	louder IF	dir_prev = up & int_prev > p4 & metr_strength > 2	747	(6.42 %)	.847	.761	.782
DL2	louder IF	mel_contour = up_down & int_prev > min3 & metr_strength > 2	890	(7.65 %)	.734	.731	.731
DL3**	louder IF	prev_dur_ratio $\leq$ 0.5 & dir_prev = up & metr_strength > 3	359	(3.09 %)	—	.709	.709
DS1	softer IF	prev_dur_ratio > 5.0	377	(4.00 %)	.764	.675	.710
DS2	softer IF	dir_prev = down & int_prev > maj3 & metr_strength $\leq$ 1 & dur_prev > 0.33	173	(1.83 %)	.745	.811	.783
DS3	softer IF	dir_prev = down & int_prev > p5 & metr_strength $\leq$ 1	169	(1.79 %)	.840	.797	.813
AS1	staccato IF	marked_staccato = yes	3,071	(13.88 %)	.916	.938	.934
AS2	staccato IF	int_next = unison	2,929	(13.23 %)	.981	.996	.934
AS3	staccato IF	int_next > p4 & dir_next = up & metr_strength $\leq$ 2	1,237	(5.59 %)	.656	.796	.756
AS4	staccato IF	next_dur_ratio $\leq$ 0.4 & dir_prev = down	1,215	(5.49 %)	.571	.809	.717
AL1	legato IF	staccato = no & mel_contour = up_down	687	(7.42 %)	.593	.513	.537

Table 2: Summary of rules discovered (*: for slow pieces only; **: for fast pieces only).

Category	#rules	True Positives		False Positives		Precision
lengthen	4	3069	(22.89%)	1234	(6.00%)	.713
shorten	2	397	(2.98%)	179	(0.87%)	.689
louder	3	1318	(11.33%)	591	(3.24%)	.690
softer	3	625	(6.63%)	230	(1.14%)	.731
staccato	4	6916	(31.25%)	1089	(5.74%)	.864
legato	1	687	(7.42%)	592	(1.86%)	.537

Table 3: Classification accuracy of learned rulesets on training data (13 Mozart sonatas).

and accuracy) of the rules on the 13 given training performances. That gives an indication of how much of the expressive variation observed in these particular performances can be 'explained' by the rules. The *generality* of the rules is then assessed by measuring their *predictive accuracy* on (a) performances of Mozart sonatas by a *different pianist*, and (b) performances of pieces of a *different musical style* (Chopin), by a variety of different pianists.

5.1 Degree of Fit on the Training Data

In trying to assess the quality of the rule sets as partial models, the first aspect we look at is the *fit* (coverage and precision) of the rule sets on the training performances they were learned from. Table 3 gives the overall coverage and precision of the induced rule sets on the training data, separately for each prediction category.

As can be seen, the categories for which rules of high coverage (and still reasonably high discriminative power) could be found are IOI lengthening and *staccato* or micro-pause insertion. The rules for dynamic attenuation (class *softer*) exhibit reasonable precision, but cover fewer cases. The other three categories turned out to be more difficult to predict at the note level. The remarkable result is that for the former three categories, such a high proportion of all observed occurrences can be predicted by so few (and simple) rules.

A striking examples of this is shown in Figure 1, which compares the pianist's timing in the first section of the A major sonata K.331 to the tempo curve predicted by the learned rules. It turns out that essentially *two (!)* simple principles/rules are sufficient to "explain" and reproduce the pianist's interpretation extremely well. All *ritardandi* (the points below the 1.0 line) but one are predicted/produced by rule TL2a (the one exception is the lengthening of the sixteenth-note D5 in bar 4, which is due to rule TL1), and all the shortenings are predicted by rule TS1.

5.2 Generality I: Testing on a Different Pianist

The purpose of the next experiment was to assess the degree of performer-specificity of the discovered rules, by testing them on performances of the same pieces, but by a different artist. The renowned pianist Philippe Entremont has also recorded some of Mozart's sonatas on a Bösendorfer SE290. We managed to obtain and process his renditions of the following pieces: Sonatas K.282 and K283 complete; plus second movements of K.279,

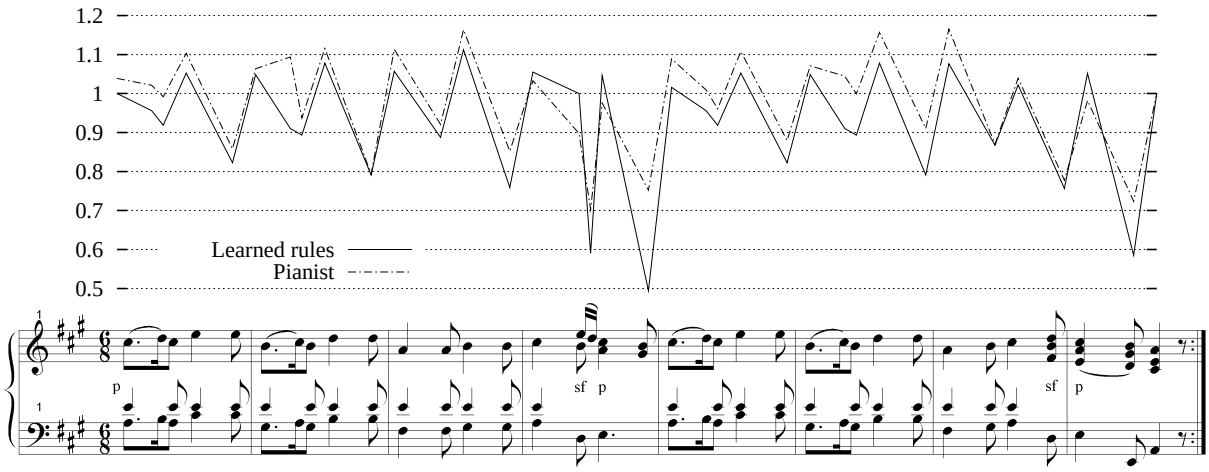

Figure 1: Timing curves for Mozart sonata K.331, first movement, first section: pianist (dashed) vs. predictions by learned rules (solid line). The precise numeric amount of lengthening/shortening was predicted by analogy from lengthenings/shortenings observed in similar situations covered by the same rule (with a *k-nearest-neighbor* algorithm with $k = 3$).

K.280, K.281, K.284, and K.333. The resulting set of performed soprano notes comprises 8,105 notes. Table 4 summarizes the predictive accuracy of our learned rules on the Entremont performances.

Comparing this to Table 3, we find no significant degradation in coverage and accuracy (except in category *softer*). On the contrary, for some categories (*lengthen*, *louder*, *staccato*) the coverage of positive examples is higher than on the original training set. The discriminative power of the rules (captured by the precision values) remains roughly at the same level. This (surprising?) result testifies to the generality of the discovered principles (and the merits of our rule discovery method).

We are currently extending the Entremont data set with recordings of three additional complete sonatas that are not present in the original training set (K.309, 310, and 311). That will provide further insight into the true generality of the rules.

Category	#rules	True Positives		False Positives		Precision
lengthen	4	596	(29.27%)	242	(7.62%)	.711
shorten	2	90	(4.10%)	45	(1.49%)	.667
louder	3	210	(13.12%)	87	(2.85%)	.707
softer	3	53	(3.32%)	45	(1.65%)	.541
staccato	4	861	(39.28%)	228	(5.71%)	.791
legato	1	131	(4.63%)	57	(1.70%)	.697

Table 4: Classification accuracy of learned rulesets on test data (Mozart performances by P.Entremont).

Category	#rules	True Positives		False Positives		Precision
lengthen	4	1752	(69.06%)	327	(10.94%)	.843
shorten	2	1472	(53.20%)	110	(4.01%)	.930
louder	3	601	(25.13%)	285	(11.06%)	.678
softer	3	0	(0.00%)	0	(0.00%)	—
staccato	4	950	(32.40%)	166	(5.92%)	.851
legato	1	17	(0.85%)	27	(0.73%)	.386

Table 5: Classification accuracy of learned rulesets on test data (performances of 2 Chopin pieces by 22 pianists).

5.3 Generality II: Testing on a Different Style

An additional experiment tested the generality of the discovered rules with respect to musical style. The rule sets were applied to two pieces by Frédéric Chopin (the first 20 bars of the Etude Op.10, No.3 in E major, and the first 45 bars of the Ballade Op.38, F major), and their predictions compared to performances (on a Bösendorfer SE290) of these pieces by 22 skilled pianists from the University of Music in Vienna. The soprano parts ('melodies') of these 44 performances amount to 6,088 notes. The coverage and precision achieved by our simple rule sets are given in Table 5.

This result is even more surprising. The categories *softer* and *legato* turn out to be basically unpredictable, and the rules for class *louder* cover a high percentage of positive examples, but also exhibit a rather high level of false predictions. But the results for the other classes (*lengthen*, *shorten*, and *staccato*) are extremely good, better in fact than on the original (Mozart) data which the rules had been learned from! The high coverage (TP) values, especially of the timing rules, are remarkable. A closer look at the Chopin test pieces shows that in the timing dimensions, three out of the six learned rules are sufficient to jointly produce these high TP rates of 69.06% and 53.20%. Both of these two test pieces have a rather regular rhythmic structure, and we plan to test the rules on a more diverse set of Chopin pieces.

Remember also that the data represent a mixture of 22 different pianists. When looking at how well the rules fit individual pianists, we find that some of them are predicted extremely well (e.g., pianist #15: timing/lengthen: TP = 89/122 (72.95%), FP = 4/129 (3.10%), π = .957; timing/shorten: TP = 71/120 (59.17%), FP = 3/132 (2.27%), π = .959).

In summary, we feel that these initial results are quite remarkable. They strongly indicate that it is possible to discover some basic performance principles (in complex, 'real-world' data) that are both fairly precise and general across a range of performers and musical styles.

6 Conclusions and Further Research

The results presented here, though promising, are only a small first step towards the ambitious goal of a comprehensive computational model of expressive performance. What

we have discovered to date are a few isolated rules that may indeed represent quite general and robust local expression principles. The rules found by the machine are very basic and simple and do in many cases conform with observations made by other researchers. What seems unique about our results is that (a) the rules have been autonomously discovered by a machine from actual performances, and (b) their validity has been tested empirically on a large number of performances by different pianists.[4] The generality of the rules across performers and styles that is suggested by our experimental results is surprising indeed, but needs more empirical validation. If the rules do turn out to be sufficiently reliable, they may form the nucleus of a more complex, multi-level model of performance.

Our next steps in this line of research are quite clear. First, we plan to perform additional experiments to evaluate the rules on different performers and different types of music. Second, we plan to extend the representation of the music with some important structural dimensions that are currently lacking, notably, *harmony*. There are some promising algorithms for automated harmony analysis (e.g., (Temperley & Sleator, 1999) that we are currently looking at. This may lead to the discovery of a few additional note-level principles. And thirdly, the next large step will be to go beyond the level of individual notes and look directly for structural performance regularities at higher levels of musical organization, in particular, *phrase structure* (Widmer, 1996). We will take a more detailed look at the empirical validity of some published models of phrase-level timing and dynamics (Todd, 1989, 1992) vis-a-vis our complex performance data. The next step will then be to study ways of combining note-level and phrase-level models (of varying degrees of abstraction) into one comprehensive model that explains as much of the observed performance patterns as possible.

7 Acknowledgments

This research is supported by the START project Y99-INF, financed by the Austrian Federal Ministry for Education, Science, and Culture, and the EU Project HPRN-CT-2000-00115 (MOSART). I would like to thank the pianists Roland Batik and Philippe Entremont for allowing us to use their performances, and the L. Bösendorfer company, Vienna, and in particular Fritz Lachnit for providing the data and technical help. Thanks to Emilios Cambouropoulos, Simon Dixon, and Werner Goebl for their help in data preparation and for helpful comments on this paper.

References

Bresin, R. (2000). *Virtual Virtuosity: Studies in Automatic Music Performance.* Doctoral Dissertation, Royal Institute of Technology (KTH), Stockholm, Sweden.

Bresin, R., and Widmer, G. (2000). Production of Staccato Articulation in Mozart Sonatas Played on a Grand Piano. Preliminary Results. *Speech Music and Hearing Quarterly Progress and Status Report* 4/2000, KTH, Stockholm.

[4] By the way, it should be mentioned at least once that preparing and documenting such a large and complex performance data set is a very laborious process and in fact involves a number of interesting research questions, such as tempo and beat induction, quantization, score-to-performance matching, etc. (Widmer, 2001a).

Cambouropoulos, E. (2000). From MIDI to Traditional Musical Notation. In *Proceedings of the AAAI'2000 Workshop on Artificial Intelligence and Music*, 17th National Conference on Artificial Intelligence (AAAI'2000), Austin, TX. Menlo Park, CA: AAAI Press.

Canazza, S., De Poli, G., Rinaldin, S., and Vidolin, A. (1997). Sonological Analysis of Clarinet Expressivity. In M. Leman (Ed.), *Music, Gestalt, and Computing.* Berlin: Springer-Verlag.

Friberg, A. (1991). Generative Rules for Music Performance: A Formal Description of a Rule System. *Computer Music Journal* 15(2), 56–71.

Friberg, A. (1995). *A Quantitative Rule System for Musical Performance.* Doctoral Dissertation, Royal Institute of Technology (KTH), Stockholm, Sweden.

Fürnkranz, J. (1999). Separate-and-Conquer Rule Learning. *Artificial Intelligence Review* 13(1), 3–54.

Gabrielsson, A. (1987). Once Again: The Theme from Mozart's Piano Sonata in A Major (K.331). In A. Gabrielsson (ed.), *Action and Perception in Rhythm and Music.* Stockholm: Royal Swedish Academy of Music.

Gabrielsson, A. (1994). Intention and Emotional Expression in Music Performance. In A. Friberg, J. Iwarsson, E. Jansson & J. Sundberg (Eds.), *Proceedings of the Stockholm Music Acoustics Conference.* Stockholm: Royal Swedish Academy of Music.

Gabrielsson, A. (1999). The Performance of Music. In D. Deutsch (Ed.), *The Psychology of Music* (2nd ed., pp. 501-602). San Diego: Academic Press.

Gabrielsson, A., and Juslin, P. N. (1996). Emotional Expression in Music Performance: Between the Performer's Intention and the Listener's Experience. *Psychology of Music* 24, 68-91.

Mitchell, T.M. (1997). *Machine Learning.* New York, NY: McGraw-Hill.

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

Palmer, C. (1988). *Timing in Skilled Piano Performance.* Ph.D. Dissertation, Cornell University.

Palmer, C. (1996). Anatomy of a Performance: Sources of Musical Expression. *Music Perception* 13(3), 433–453.

Parncutt, R. (1993). Categorical Perception of Short Rhythmic Events. *Proceedings of SMAC'93*, Royal Swedish Academy of Music.

Repp, B. (1992). Diversity and Commonality in Music Performance: An Analysis of Timing Microstructure in Schumann's 'Träumerei'. *Journal of the Acoustical Society of America* 92(5), 2546–2568.

Repp, B. H. (1998). A Microcosm of Musical Expression: I. Quantitative Analysis of Pianists' Timing in the Initial Measures of Chopin's Etude in E major. *Journal of the Acoustical Society of America* 104, 1085–1100.

Repp, B. H. (1999). A Microcosm of Musical Expression: II. Quantitative Analysis of

Pianists' Dynamics in the Initial Measures of Chopin's Etude in E major. *Journal of the Acoustical Society of America* 105, 1972–1988.

Seashore, C.E. (ed.) (1936). *Objective Analysis of Music Performance.* Iowa City, IA: University of Iowa Press.

Shaffer, L.H. (1980). Analyzing Piano Performance: A Study of Concert Pianists. In G.Stelnmach and J. Requin (eds.), *Tutorials in Motor Behavior.* Amsterdam: North-Holland.

Shaffer, L.H., Clarke, E., and Todd, N. (1985). Metre and Rhythm in Piano Playing. *Cognition* 20, pp.61–77.

Sundberg, J. (1993). How Can Music Be Expressive? *Speech Communication* 13, 239–253.

Sundberg, J., Friberg, A., and Frydén, L. (1991). Common Secrets of Musicians and Listeners: An Analysis-by-Synthesis Study of Musical Performance. In P. Howell, R. West & I. Cross (eds.), *Representing Musical Structure.* London: Academic Press.

Temperley, D. and Sleator, D. (1999). Modeling Meter and Harmony: A Preference Rule Approach. *Computer Music Journal* 23(1), 10–27.

Timmers, R., Ashley, R, Desain, P, and Heijink, H. (2000). The Influence of Musical Context on Tempo Rubato. *Journal of New Music Research* 131–158.

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

Todd, N. (1992). The Dynamics of Dynamics: A Model of Musical Expression. *Journal of the Acoustical Society of America* 91, pp.3540–3550.

Widmer, G. (1996). Learning Expressive Performance: The Structure-Level Approach. *Journal of New Music Research* 25(2), 179–205.

Widmer, G. (2000). Large-scale Induction of Expressive Performance Rules: First Quantitative Results. In *Proceedings of the International Computer Music Conference (ICMC'2000).* San Francisco, CA: International Computer Music Association.

Widmer, G. (2001a). Using AI and Machine Learning to Study Expressive Music Performance: Project Survey and First Report. *AI Communications* 14(3), 149–162.

Widmer, G. (2001b). Discovering Strong Principles of Expressive Music Performance with the PLCG Rule Learning Strategy. In *Proceedings of the 11th European Conference on Machine Learning (ECML'01)*, Freiburg. Berlin: Springer Verlag.

Windsor, L. and Clarke, E. (1997). Expressive Timing and Dynamics in Real and Artificial Musical Performances: Using an Algorithm as an Analytical Tool. *Music Perception* 15, 127–152.

Windsor, L., Desain, P., Honing, H., Aarts, R., Heijink, H., and Timmers, R. (2000). On Time: The Influence of Tempo, Structure and Style on the Timing of Grace Notes in Skilled Musical Performance. In Desain, P. and Windsor, W. L. (Eds.), *Rhythm Perception and Production.* Lisse: Swets & Zeitlinger.

Witten, I.H. & Frank, E. (1999). *Data Mining.* San Francisco, CA: Morgan Kaufmann.