Using AI and Machine Learning to Study Expressive Music Performance: Project Survey and First Report

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

This article presents a long-term inter-disciplinary research project situated at the intersection of the scientific disciplines of Musicology and Artificial Intelligence. The goal is to develop AI, and in particular machine learning and data mining, methods to study the complex phenomenon of *expressive music performance.* Formulating formal, quantitative models of expressive performance is one of the big open research problems in contemporary (empirical and cognitive) musicology. Our project develops a new direction in this field: we use *inductive learning techniques* to discover general and valid expression principles from (large amounts of) real performance data. The project is currently starting its third year and is planned to continue for at least four more years.

In the following, we explain the basic notions of expressive music performance, and why this is such a central phenomenon in music. We present the general research framework of the project, and discuss the various challenges and research opportunities that emerge in this framework. We then briefly describe the current state of the project and list the main achievements made so far. In the rest of the paper, we discuss in more detail one particular data mining approach (including a new algorithm for learning characterisation rules) that we have developed recently. Preliminary experimental results demonstrate that this algorithm can discover very general and robust expression principles, some of which actually constitute novel discoveries from a musicological viewpoint.

Keywords: Machine Learning, Data Mining, Expressive Music Performance

Draft version of a paper to appear in
AI Communications 14, IOS Press (2001).

1. Introduction

The use of Artificial Intelligence methods, particularly inductive machine learning, as discovery tools in scientific research has found increasing interest in recent years. Especially in the biosciences (biochemistry, genetics, etc.) where there are large amounts of empirical data of highly complex structure, there have been numerous attempts at using machine learning and data mining for intelligent data analysis and scientific discovery (see, e.g., [11,24,25,27,39]). Less well investigated is the potential of AI-based data analysis and discovery methods in the so-called humanities or 'softer' sciences, which at first sight would seem to defy attempts at machine-based discovery because of a lack of strict laws that can be discovered. The project to be described here falls into this latter category. We apply AI methods to perform research in *empirical musicology.*

More precisely, the project aims at developing computational methods to study the phenomenon of *Expressive Music Performance* (or *'Musical Expression'*) and to inductively build formal models of (particular aspects of) expressive performance from real performances by human musicians. The emphasis here is on the word *inductively*: what distinguishes our approach from most of the other work in empirical musicology is that we want to use large amounts of 'real-world' performance data as a basis for our investigations, so that the resulting hypotheses and models have a strong empirical foundation. To deal with the complexity of such large amounts of data, we need the help of intelligent data analysis methods; hence the use of machine learning and data mining. The hope is that such methods will enable us to discover interesting and possibly novel patterns and regularities in the data, and to specify models of expressive performance in a formal way that makes possible extensive and precise testing.

Formulating quantitative models of expressive performance is one of the big open research problems in contemporary (empirical and cognitive) musicology. Our approach is meant to complement other researchers' work in performance research (e.g., [9,29,32,34,35,36,37,38,40,42,43], which has produced a wealth of detailed hypotheses and insights, but has often been based on rather limited sets of performance data (which were sometimes also produced under 'laboratory conditions'). In short, we aim at performing the most data-intensive investigations ever done in musical expression research.

The project, financed by a generous research grant by the Austrian Federal Government (see Acknowledgements), started in early 1999 and is intended to last six years. The research is truly inter-disciplinary, involving both musicologists and AI researchers. It also contributes new results to both disciplines involved. (For instance, the research described in section 5 has produced both new insight into common performance patterns and a new, general data mining algorithm.)

The purpose of the present article is to give an overview of the project goals, to describe the current project status, and to present first results that show that the chosen approach is viable. The article is organized as follows: in section 2, we explain the basic notions of expressive music performance, and why this is such a central phenomenon in music. Section 3 presents the general research framework of the project, and discusses the various challenges and research opportunities that emerge in this framework. Section 4 briefly describes the current state of the project and lists some of the main technical achievements made so far. In section 5 we then take a more detailed look at one particular data mining approach (including a new algorithm for learning characterization rules) that we have developed just recently. Preliminary experimental results demonstrate that this algorithm can discover very general and robust expression principles, some of which actually constitute novel discoveries from a musicological viewpoint. The article concludes with a brief discussion of potential applications of our methods (section 6), and a preview of planned research directions for the coming years (section 7).

2. Expressive Music Performance

When played exactly as notated in the musical score, a piece of music would sound utterly mechanical and lifeless; it is both unmusical and physically impossible for a musician to perform a piece with perfectly constant tempo, even loudness, etc. What makes a piece of music come alive (and what makes some performers famous) is the art of *music interpretation*, that is, the artist's understanding of the structure and 'meaning' of a piece of music, and his/her (conscious or unconscious) expression of this understanding via *expressive performance*: a performer shapes a piece by continuously varying important parameters like tempo, dynamics (loudness), articulation, etc., speeding up at some places, slowing down at others, stressing certain notes or passages by various means, and so on. What types of parameters are at a performer's disposal partly depends on the instrument being played (consider, e.g., vibrato). It is this shaping that can turn a lifeless piece of music into a moving experience, and that also makes both the composer's and the performer's ideas clear to the listener.

It is important to realize that expressive variation is more than just a 'distortion' of the original (notated) piece of music. In fact, the opposite is the case: the notated music score is but a small part of the actual music. Not every intended nuance can be captured in a limited formalism such as common music notation, and the composers were and are well aware of this. The performing artist is an indispensable part of the system, and expressive music performance thus plays a central role in our musical culture.

This has been reflected in musicological research only gradually.[1] While in the past, the vast majority of music research dealt with the analysis of musical works, their structure, and formal theories thereof, recent work in (empirical) musicology is now increasingly focusing on aspects of expressive performance. The main questions to be answered

[1] One reason for this may have been the difficulty of obtaining reliable measurements, which, before the advent of MIDI instruments, required the construction of complex (and often imperfect) measuring devices. In fact, extracting the details of real performances from audio recordings is still a big problem and has prompted us to invest quite some research effort into the development of music (MIDI and audio) analysis algorithms — see section 4.2 below.

are: are there explainable and quantifiable principles that govern expressive performance? to what extent and how are 'acceptable' performances determined by the (structure of the) music? what are the cognitive principles that govern the production (in the performer) and the perception (in the listener) of expressive performances? and what does this have to do with how we experience music?

The project described here hopes to contribute to answering the first two of these questions. We collect data representing precise measurements of expressive performances by skilled musicians, and try to detect patterns and regularities (and intelligible characterizations of these) with the help of machine learning and data mining methods. As we also enable the computer to recognize structural aspects of the music, potential relationships between expressive patterns and musical structure should emerge naturally from these investigations.

This approach is based on earlier work by the author [46,47,48,49], where it was shown that given some knowledge about musical structure, a computer can indeed learn general performance rules that produce rather sensible 'interpretations' of musical pieces. The central problem with these early studies was a lack of real performance data (the investigations were based largely on performances by the author himself). A major goal of the current project is to go beyond this by working with large collections of performances by skilled musicians, preferably recorded under 'natural' conditions (i.e., not in the laboratory, with the experimental purpose in mind). Ideally, we would also like to study the performance style of famous artists, but that will depend on the availability of computational methods for precise musical information extraction from audio, which is still an open research problem (see section 4.2 below).

The main dimensions of expressive variations that are the focus of our current studies are *tempo* and *timing*, *dynamics* (loudness variations), and *articulation* (the use of overlaps or pauses between successive notes). With regard to the music, we restrict our attention to *tonal music*, as practiced in our part of the world since the 17th century. At the moment we are working exclusively with classical piano music, but the studies may also be extended to other instruments and other (tonal) styles (such as jazz).

3. Inducing Performance Models from Real Performances: The Basic Research Framework

Our goal is to inductively build models of expressive performance from measurements of real performances by human musicians. An example might be a predictive model of tempo changes in the form of a set of classification and regression rules. To achieve this, it is necessary to

1. obtain high-quality performances by human musicians (e.g., pianists) in machine-readable form;
2. encode the notated scores of the corresponding pieces in machine-readable form;
3. extract the 'expressive' aspects (e.g., tempo and loudness deviations) from these (by comparing notated scores and actual performances) and transform these into data that are amenable to computer analysis (e.g., tempo and dynamics curves);
4. analyze the structure (meter, grouping, harmony, etc.) of the pieces and represent the scores and their structure in a formal representation language;
5. develop machine learning algorithms that search for systematic connections between structural aspects of the music and typical expression patterns, and formulate their findings as symbolic rules;
6. perform systematic experiments with different representations, sets of performances, musical styles, etc.; and
7. analyze the learning results with a view to both qualitative (are the discovered rules musically sensible? interesting? related to theories by other researchers?) and quantitative aspects (how much of the variance can be explained? where are the limits?).

Figure 1 sketches the research framework, the data collections involved, and the necessary data processing and analysis steps. Each of these steps turns out to be highly non-trivial and presents us with a number of problems, or, better, challenging research opportunities.

The task of *preparing and documenting performance data*, which at first sight seems rather straightforward and boring (though important), actually turns out to be difficult and extremely work-intensive, given the massive amounts of data we succeeded in acquiring (see section 4.1). For in-

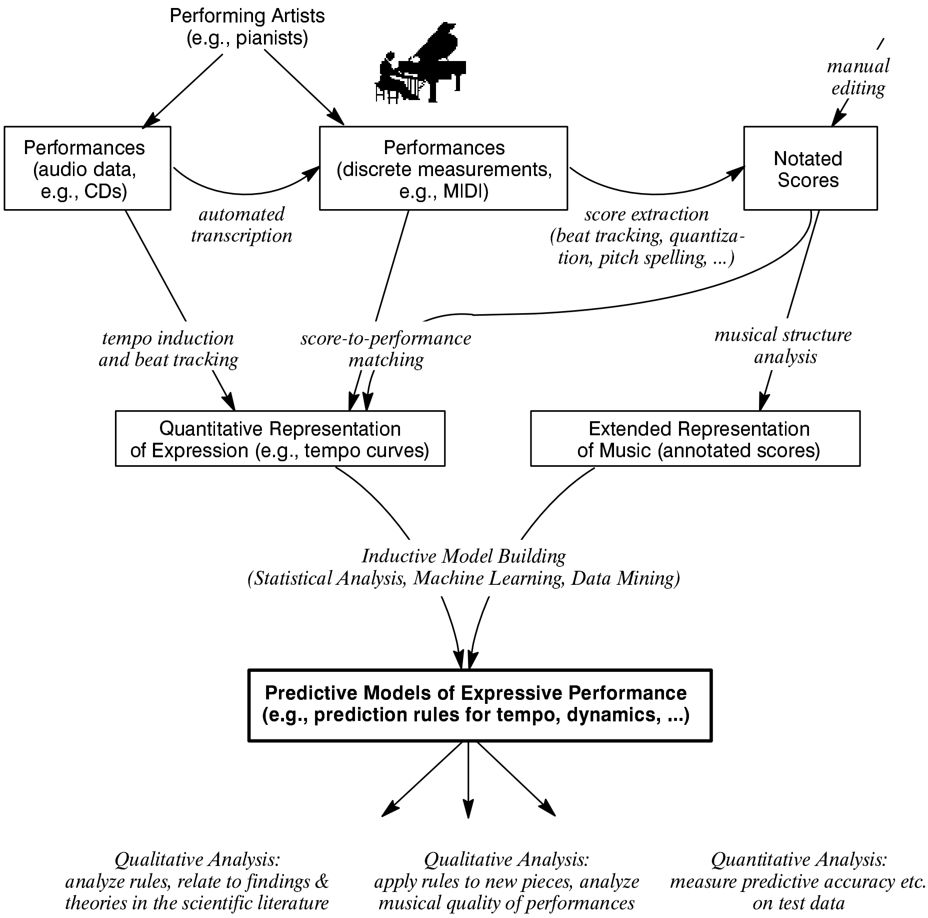

Fig. 1. The research framework: a sketch of data and processing/analysis steps.

stance, coding musical scores consisting of tens of thousands of notes by hand is not feasible; in order to get at the scores, we needed to develop computational methods for extracting (re-constructing) score information from the performances themselves. The result are a number of novel algorithms of general utility (see section 4.2).

Likewise, *musical structure analysis* is an open research problem. Structural aspects of the music, such as harmony, metrical structure, phrase structure, etc. are assumed to be important for explaining expressive choices and must thus be included in the representation of the training pieces. Again, a complete manual analysis of a large number of complex pieces is infeasible or at least highly impractical, so there is a need for computational methods. In the context of our project, a number of intelligent analysis algorithms have been developed (see section 4.3), but much more remains to be done. In particular, there are currently no reliable algorithms for harmonic analysis and phrase structure analysis, both of which would have a large number of practical applications outside our project. This is a veritable challenge to AI research.

The challenges in the domain of *machine learning* and *data mining* are manifold. The *representation* issue, relevant in any inductive learning scenario, is of central importance in our project. There are many conceptual frameworks in which music can be described, and the choice of representation vocabulary crucially affects what can be learned. Systematic experimentation will be necessary to identify those dimensions of musical structure and those musical concepts that are most relevant to explaining and predicting observed expression patterns. A related question concerns the appropriate *abstraction level*, not only for the representation of the music, but also concerning the definition of the target concepts themselves: what types of patterns are we looking for, and at what musical abstraction levels? For instance, some aspects of performance may be describable at the level of single notes (e.g., the lengthening of a particular note to delay an important event — see section 5 below), but other observed phenomena may only make sense when viewed as parts of larger, more abstract patterns (e.g., a note lengthening as part of a large-scale *ritardando* towards the end of a phrase or section). In the end, it will be necessary to learn models at different structural abstraction levels, which leads us to the problems of discerning and separating multiple pattern levels in given training observations, and of combining learned models of different granularity. These questions definitely bear relevance to machine learning and data mining beyond this particular project.

And finally, there is the *evaluation* issue. How is one to evaluate and quantify the degree of validity of a given model in a domain where there is no unique 'correct' solution (there are usually many 'acceptable' ways of performing a piece), and where it is clear from the outset that there can be no model that even comes close to 100 % prediction accuracy (after all, we cannot assume that an artist's complex choices and behaviour are perfectly predictable)? The empirical evaluation methods used in machine learning (measuring classification accuracy and prediction error on unseen data, estimating true error via cross-validation etc.) do have their place in this research, but they need to be complemented with more music-specific methods that, while avoiding to make judgments concerning the musical or aesthetic quality of a performance, do account for musical aspects of a model's predictions.

All in all, the ambitious goal of the project forces us to tackle a broad range of problems in the intersection of Artificial Intelligence and Music. This is also reflected in the diversity and breadth of the research currently pursued within the project, as briefly described in the following section.

4. A Brief Activity and Status Report

In the following, we briefly summarize the main research activities and achievements of the first two project years. This section may be skipped by readers only interested in the machine learning aspects of the project.

4.1. 'Real-world' Performance Data

At the moment, we are working with piano performances. As it is impossible, with current mathematical and computational methods, to extract precise performance information (start and end times, loudness, etc. of each individual note) directly from audio data such as is available from records and CDs (though we are working on solutions to some specialized problems — see 4.2), the main source of performance data are special pianos that precisely record each performer action. For example, the Bösendorfer SE290 is a full concert grand piano with a special mechanism that measures and records every key and pedal movement with high precision. (The piano also features a mechanical reproduction facility that can re-produce a recorded performance with very high accuracy.) Several high-class pianists have made recordings on this piano, and we have managed to obtain the permission to use several large sets of such performances. These include (a) 13 complete Mozart piano sonatas performed by a Viennese pianist, (b) 5 complete and 6 partial Mozart piano sonatas by a French pianist, (c) a set of 22 x 2 Chopin performances (2 pieces played by 22 Viennese pianists), and (d) essentially the entire piano works by Frédéric Chopin, played by a famous Russian pianist. To give the reader an impression of the complexity of these data sets: the 13 Mozart sonatas (a) alone amount to some 3.5 Megabytes of data representing some four hours of music, more than 106,000 notes, and 500,000 measurements of pedal action. The Chopin corpus (d) (not pre-processed yet — see below) comprises more than 2 million events. Thus, we are working with what is most probably the largest fully measured collection of real performances ever compiled in the history of performance research.

4.2. Score and Expression Extraction

Preprocessing these data to make them usable for analysis and machine learning is a formidable task. What we need is not only the performances (i.e., information about how the notes were played), but also the notated score (i.e., information about how the notes 'should be' played) and the correspondence between each score note and each played note, including information about which notes were omitted or missed, and which extra notes were played by the pianist, unintentionally or as part of ornaments. As typing in all the score information by hand was clearly out of the question, we decided to try to extract the relevant score information directly from the (expressive) performance. Score extraction involves a number of non-trivial problems. The main problem is identifying a regular pattern of beats in a performance 'distorted' by expressive tempo changes. Consequently, we had to develop new algorithms for

- *beat induction*, including tempo identification and the tracking of tempo changes [13,15,17]; that involves finding the most salient level of beats (the 'tactus') and being able to track the positions of strong beats even in situations where the tempo changes significantly;
- *quantization*, i.e., inferring the 'intended' onset times and duration of notes in the underlying score [3]; especially inferring the intended duration is difficult and only partly possible, given the wild mixtures of staccato and legato of varying degrees applied by the pianists;
- inferring the correct or intended *enharmonic spelling* of notes (e.g., G♯ vs. A♭) [4], which is not merely an aesthetic issue, but absolutely vital for the correct interpretation of a musical passage.

The 'raw' score files extracted by these algorithms (tens of thousands of lines of text) were then manually corrected and further annotated by a number of students. In a semi-automatic process, the score files were then matched onto the performance files to establish the exact note-to-note correspondence; thousands of notes were manually identified and labelled as missing or extraneous (most of these are related to ornaments like trills etc.).

Another line of research concerns the extraction of performance information directly from *digital audio* data (e.g., CDs), which would then allow us to also study arbitrary recordings by famous artists [14]. Our current methods have been integrated into an interactive tempo analysis tool [16] and are currently used to measure and classify the beat-level tempo behaviour of several famous pianists ([22]).

4.3. Musical Structure Analysis

One of our guiding hypotheses (which is in accordance with the opinion of other researchers in this area) is that one purpose of expressive music performance is for the performer to communicate his/her interpretation of the structure of a piece of music to the listener, and in fact to make the musical structure audible. Thus, if we want the computer to discover and 'explain' common performance patterns, it must be capable of identifying relevant structural units in a given piece of music. In the first phase of the project, three aspects of algorithmic music analysis were investigated:

- *Segmentation*: A computational model of automatic music segmentation was developed and tested on a variety of musical styles. The goal is to decompose a piece into contiguous segments such as motifs, groups, phrases, etc. That is one of the most evident dimensions along which human listeners structure a heard piece, and which performers emphasize and sometimes disambiguate by expressive 'phrasing'. Using this model, we found some interesting correlations between musical segment boundaries and a pianist's typical timing behaviour [6].
- *Categorization and motivic analysis*: We have developed an algorithm that attempts to group musical segments into meaningful categories or clusters, that is, it tries to find groups of melodic motifs [7]. At the heart of the algorithm is a new, general-purpose clustering method. In experiments, the algorithm succeeded in partly reproducing motivic analyses by human musicologists of such complex pieces as Schumann's *Träumerei* and Debussy's *Syrinx* [8]. It was also demonstrated experimentally that our model performs well in predicting the categorizations made by humans listeners [5].
- *Implication-Realization Model*: We have implemented selected parts (the lowest level) of E. Narmour's *Implication-Realization Model* [28], which identifies various types of melodic and rhythmic patterns that may be meaningful units in the perception and production of music.

The analyses computed by these tools can be used as descriptors in the representation of musical pieces. Two important aspects that are currently missing are *harmony* and *phrase structure*. Despite some good recent work in this area [41], developing reliable and musically intelligent algorithms for these problems remains an open challenge for AI research.

4.4. Systematic Musical and Statistical Analyses

A number of specific aspects of expressive performance have been or are currently being studied on the basis of our data sets. The following items are not so interesting from an AI point of view, but we briefly include them here for the sake of completeness.

- *Performance averaging*: Using the 2 x 22 Chopin performances, we showed that even though there are interesting differences between the individual pianists, a 'generic' performance synthesized by averaging the 22 individual performances produces a surprisingly well-sounding musical result. This corroborates recent results by Bruno Repp, of Yale University [33], and points to interesting possibilities for further research on common performance standards [20].
- *Melody lead*: Another study on the same data set concerned the phenomenon of 'melody lead', i.e., the fact that the notes of the melody are generally played slightly ahead of the other voices (by 10 to as much as 50 milliseconds). We found strong evidence that this phenomenon is probably not really used intentionally as an expressive device by pianists (as hypothesized by some researchers [30]), but rather that it is just an artifact caused by mechanical properties of the piano [21].
- *Articulation*: In a collaboration with a research group at the Royal Institute of Technology (KTH) in Stockholm, we discovered some interesting dependencies between a pianist's use of *staccato* articulation and musical context [2], and these findings have led to the inclusion of several new rules in KTH's well-known performance rule set [1].

4.5. Model Building via Inductive Learning: Initial Investigations

In the domain of machine learning and data mining proper, initial experiments concerned questions of general learnability. In one set of experiments, we attempted to learn parameters that optimally fit an existing rule-based model of expressive performance (the KTH rule set [18]) to our large set of Mozart data, with the help of a genetic algorithm. The preliminary experimental results [26] are surprising in that they seem to indicate that — at least given the fitness and quality measures we used — there are no settings of the rule parameters that consistently produce results better than the 'baseline' (not applying any rules at all) over a large number of pieces, though good parameter settings could be found for individual pieces. This finding needs further investigation.

A second initial study investigated the general feasibility of inducing new classification rules at the most basic musical level: the level of individual notes. The goal is to discover rules that predict how individual notes will most likely be played by a pianist (e.g., louder or softer than their predecessor). In a first suite of experiments [50], we succeeded in showing that even at that low level, there is substantial structure in the data that can be revealed by an inductive learning algorithm; the learning algorithms were able to find rule sets that predict the performer's choices with better than chance probability.

However, the improvement over the baseline accuracy was generally rather small (though statistically significant), which indicates that there are severe limits as to how much of a performer's behaviour can be explained at the note level. It is clearly unreasonable to expect that one can find a note-level model that completely discriminates between different categories of performer actions. Moreover, 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 3037 leaves (after pruning)! This is clearly not desirable as the purpose of the project is to discover intelligible rules that provide new insight.

These results prompted us to develop a new approach to learning *partial* rule-based models. The goal is to induce local models for those aspects of expression that *can* be described and 'explained' at the note level, and to separate these from other effects that may require us to look at higher levels of musical structure, such as grouping and phrase structure. We will take a closer look at that in the following section.

5. Learning Partial Characterizing Models: A New Data Mining Approach

This section presents in some detail a new data mining method for discovering partial characterizing models, and presents experimental evidence that the corresponding algorithm can discover very general and robust expression principles, some of which actually constitute novel, musically interesting discoveries.

5.1. The Goal: Learning Partial Models

Starting from the assumption that the level of individual notes is not a sufficient basis for perfectly discriminating classification models, we propose to abandon the notion of complete coverage/discrimination altogether and instead to search for 'partial' rule models that 'only explain what can be explained' at the note level. The idea is to look for small numbers of 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). Such rule sets will be called *Partial Characterizing Models.*

In the following, it will be shown that with a new data mining strategy, very simple partial models can be found that still explain a number of interesting sub-categories of expressive performance behaviour. (Indeed, we will show that 4 simple rules are sufficient to predict 22.89% of the instances of note lengthening in our large data set, which contrasts nicely with the decision tree with 3037 leaves mentioned above.)

5.2. Data and Target Concepts

The data used in this study consist 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 the Viennese pianist Roland Batik on a Bösendorfer SE290 computer-monitored concert grand piano. The piano measurements (hammer impact times and pedal movements), together with the notated score in machine-readable form, provide all the information needed to compute expressive variations in timing, dynamics, and articulation.

The experiments were carried out on the melodies (mostly the soprano parts) only, which gives an effective training set of 41,116 played notes. Each note is 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).

With respect to expressive performance, the dimensions we are looking at are (local) timing, dynamics, and articulation. In trying to learn expression rules, 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 timing dimension, a note N is assigned to class *lengthen* if the inter-onset interval (IOI) between the start of N and the next note is lengthened relative to (a) its predecessor and (b) the current 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 (from note onset to note offset) 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 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 these definitions, there will be some notes that are neither examples nor counter-examples of some concept; there are some good musical reasons for this.)

5.3. The PLCG Rule Discovery Algorithm

From the perspective of data mining, the problem is to find partial models that are *characterizing* in the sense of compactly describing various sub-classes of situations that are treated in a similar way by the performer (such as 'situations where a note is lengthened' vs. 'situations where a note is shortened'); generalization accuracy on other pieces or performers is a secondary issue, but will also be evaluated in the experiments. *Partial,* as explained above, means that we do not expect the models to be able to cover and describe all of

Given:

- a set of training instances D
- a target concept (class) c
- a rule learning algorithm L
- a rule selection criterion C

Algorithm:

1. Separate the training examples D into n subsets D_i, $i = 1 \ldots n$ (randomly or according to a particular scheme);
2. Learn partial rule models $R_i = \{r_{ij}\}$ for class c from each of these subsets D_i separately, using the learning algorithm L. The resulting sets of rules R_i will most likely be overly specific.
3. Merge the rule sets R_i into one large set R: $R = \bigcup R_i$.
4. Perform a hierarchical clustering of the rules in R into a tree of clusters C_i, $i = 1 \ldots k$ of similar rules, using a standard hierarchical agglomerative clustering algorithm [23] and an appropriate syntactic/semantic rule similarity measure.
5. For each cluster C_i, compute the *least general generalization* of all the rules in C_i: $\hat{r}_i = lgg(\{r_{ij} | r_{ij} \in C_i\})$. The resulting tree T of rules $\hat{r}_i$ represents generalizations of various degrees of the original rules.
6. From this generalization tree T, select those rules r_i that optimize the given selection criterion C.

Fig. 2. Sketch of the PLCG (**P**artition+**L**earn+**C**luster+**G**eneralize) strategy for learning partial characterization rules.

(or even a large part of) the instances of a given category observed in the data. Moreover, given the nature of our target phenomenon, we cannot expect very high levels of discriminative accuracy.

These requirements suggest the use of rule learning algorithms of the *set covering* variety (also known in machine learning as *separate-and-conquer* learners [19]), such as FOIL [31] or RIPPER [10]. These algorithms learn models one rule at a time, in each rule refinement step selecting a literal that maximizes some measure of discrimination (e.g., information gain). A rule is refined (specialized) until a given stopping criterion is satisfied (typically based on the rule's purity or precision), and the overall learning process stops when no more rules can be found that satisfy this purity criterion. The stopping criterion is thus the natural entry point for the user to influence the generality and precision of the induced rules. In the context of our problem, we would require rather low levels of precision. The degree of coverage of the resulting rules would then follow automatically, dictated by the data.

We have chosen to pursue a more complex approach. The basic idea is to learn several models in parallel (from subsets of the data), search for groups of similar rules in these models, generalize these into summarizing rules, and then select those generalizations for the final model that optimize some (possibly global) user-defined criterion (which will typically be a trade-off function between *coverage* (the number of cases covered by the rule) and *precision* (the proportion of correct predictions made by the rule)). This strategy gives us more direct control over the overall coverage and precision of the induced models, and at the same time helps ameliorate one of the major problems of the greedy literal selection strategy used by the underlying rule learner: the danger of selecting sub-optimal conditions due to the local maximization of a given discrimination measure. In this sense, our approach — let us call it the PLCG (**P**artition+**L**earn+**C**luster+**G**eneralize) strategy — is inspired by the success of so-called *ensemble methods* in machine learning (see [12] for a good overview). The corresponding algorithm is given in more detail in figure 2.

PLCG is really a meta-algorithm that can be wrapped around any algorithm that learns classification rules. We are using our own implementation of a propositional FOIL-type [31] algorithm, with the standard information gain heuristic and with a parameterizable stopping criterion based on rule purity and mininum required rule coverage. As for the *rule selection criterion* (step 6 of the algorithm), we currently use a very simple one: given a minimum precision MP and a minimum coverage MC specified by the user, descend along each branch of the tree and return the rule with maximum coverage that has a precision $\geq MP$ (note that precision values grow monotonically as we descend down the tree) and a coverage $\geq MC$. Subtrees below nodes with a coverage $< MC$ can be pruned. This is a very simple criterion that looks at each rule in isolation. More complex, global criteria are easily conceivable (e.g., search for a set of rules that together maximize coverage, given a particular minimum precision — another optimization (set covering) problem).

5.4. Some Simple Principles Discovered

The PLCG algorithm was run 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 with $MC = .02$ and $MP = .7$ (from a total of 383 specialized rules) consist of 6 rules for local timing, 6 rules for local dynamics, and 5 rules for articulation. (Two rules were selected manually for musical interest, although they did not quite reach the required coverage and precision, respectively.) We lack the space to list all the rules here (see [51]). Let us illustrate the types of principles found by looking at just one of the discovered rules, as represented in our specific rule language:

RULE TL2:
```
abstract_duration_context = equal-longer
& metr_strength ≤ 1
⇒ lengthen
```

"Given two notes of equal duration followed by a longer note, lengthen the note that precedes the final, longer one, if this note is in a metrically weak position ('metrical strength' ≤ 1)."

This is an extremely simple principle that turns out to be surprisingly general and quite precise: rule TL2 correctly predicts 1,894 cases of local note lengthening, which is 14.12% of all the instances of lengthening observed in the training data. The number of incorrect predictions is 588 (2.86% of all the counterexamples). Together with a second, similar rule relating to the same type of phenomenon, rule TL2 covers 2,964 of the positive examples of note lengthening in our performance data set, which is more than one fifth (22.11%)! It is highly remarkable that one such simple principle is sufficient to predict such a large proportion of observed note lengthenings in a complex corpus such as Mozart sonatas. This is a truly novel (and surprising) discovery; none of the existing theories of expressive performance consider this simple, but general and robust, principle.

A few other interesting rules were discovered, such as two pairs of timing and articulation rules that nicely characterize the pianist's consistent treatment of certain types of melodic leaps and rhythmic patterns. There are some relationships between these discoveries and findings and hypotheses in the musicological literature, which will form the starting point for new, more specialized musicological investigations [51].

5.5. Quantitative Evaluation

To get a first idea of how well the induced models capture the pianist's performance style, we measure the *fit* (coverage and precision) of the three rule sets on the *training data*, separately for each prediction category. Table 1 gives the results.[2]

As can be seen, the classes for which rules of high coverage (and still reasonably high discriminative power) could be found are note lengthening and staccato. 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.

The next question concerns the *generality* of the discovered rules. How well do they transfer to other pieces and other performers? To assess the degree of performer-specificity of the rules, we test them on performances of the same pieces, but by a different artist. The test pieces in this case are the Mozart sonatas K.282, K283 (complete) and K.279, K.280, K.281, K.284, and K.333 (second movements), performed by the renowned conductor and pianist Philippe Entremont, again on a Bösendorfer SE290 piano. The results are given in Table 2.

[2] *True positives* = TP = number of positive examples correctly predicted; *false positives* = FP = number of incorrect predictions; *precision* = $p = TP/(TP+FP)$ = proportion of predictions that are correct.

Category	#rules	True Positives		False Positives		Precision
lengthen	4	3069/13410	(22.89 %)	1234/20551	(6.00 %)	.713
shorten	2	397/13307	(2.98 %)	179/20550	(0.87 %)	.689
louder	3	1318/11629	(11.33 %)	591/18260	(3.24 %)	.690
softer	3	625/9429	(6.63 %)	230/20113	(1.14 %)	.731
staccato	4	6916/22132	(31.25 %)	1089/18984	(5.74 %)	.864
legato	1	687/9256	(7.42 %)	592/31860	(1.86 %)	.537

Table 1
Classification performance of learned rulesets on training data (13 Mozart sonatas); percentages are relative to total number of positive and negative examples, respectively.

Category	#rules	True Positives		False Positives		Precision
lengthen	4	596/2036	(29.27 %)	242/3175	(7.62 %)	.711
shorten	2	90/2193	(4.10 %)	45/3013	(1.49 %)	.667
louder	3	210/1601	(13.12 %)	87/3055	(2.85 %)	.707
softer	3	53/1598	(3.32 %)	45/2725	(1.65 %)	.541
staccato	4	861/2192	(39.28 %)	228/3996	(5.71 %)	.791
legato	1	131/2827	(4.63 %)	57/3361	(1.70 %)	.697

Table 2
Classification performance of learned rulesets on test data (Mozart performances by P.Entremont).

Category	#rules	True Positives		False Positives		Precision
lengthen	4	1752/2537	(69.06 %)	327/2988	(10.94 %)	.843
shorten	2	1472/2767	(53.20 %)	110/2746	(4.01 %)	.930
louder	3	601/2392	(25.13 %)	285/2578	(11.06 %)	.678
softer	3	0/2249	(0.00 %)	0/2784	(0.00 %)	—
staccato	4	950/2932	(32.40 %)	166/2802	(5.92 %)	.851
legato	1	17/2011	(0.85 %)	27/3723	(0.73 %)	.386

Table 3
Classification performance of learned rulesets on test data (performances of 2 Chopin pieces by 22 pianists).

Comparing this to Table 1, we find no significant degradation in coverage and precision (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 enable us to estimate the rules' *generalization accuracy* on unseen data.

Another experiment tested the generality of the discovered rules with respect to *musical style*. They were applied to pieces of a very different style (Romantic piano music), specifically, 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 the results were compared to performances of these pieces by 22 skilled pianists. The soprano parts ('melodies') of these 44 performances amount to 6,088 notes. The coverage and precision values achieved on this data set are given in Table 3.

This result is even more surprising. The categories *softer* and *legato* turn out to be basically

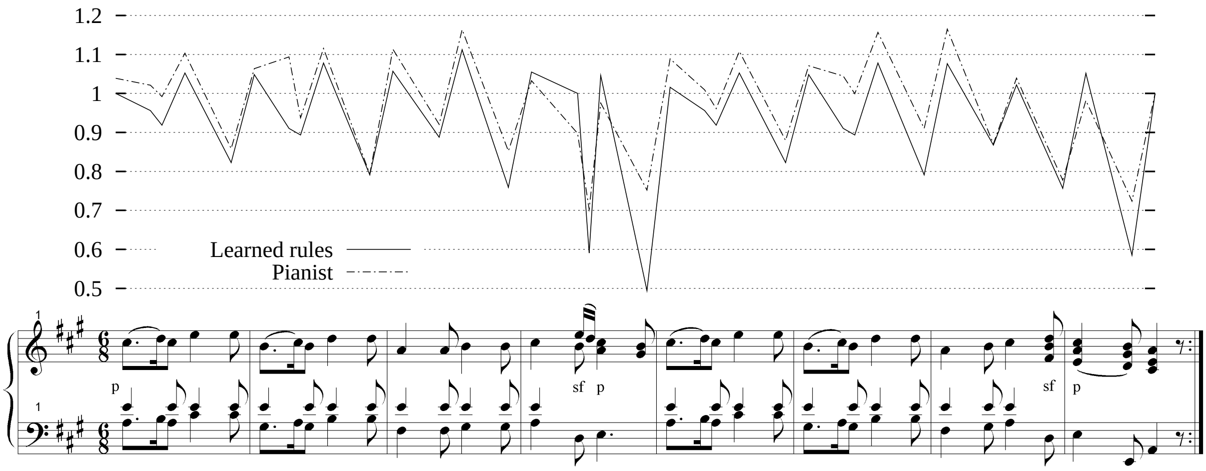

Fig. 3. Mozart Sonata K.331, 1st movement, 1st part, as played by pianist and learner (tempo). The curve plots the relative tempo at each note — notes above the 1.0 line are shortened relative to the tempo of the piece, notes below 1.0 are lengthened. A perfectly regular performance with no timing deviations would correspond to a straight line at $y = 1.0$

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%), $p = .957$; timing/shorten: $TP = 71/120$ (59.17%), $FP = 3/132$ (2.27%), $p = .959$).

Finally, to give the reader an impression of just how well a few simple rules can predict a pianist's behavior in certain cases, Figure 3 compares the tempo variations predicted by our rules to the pianist's actual timing in a performance of the well-known Mozart Sonata K.331 in A major (first movement, first section). In fact, it is just two simple rules (one for note lengthening, one for shortening) that produce the system's timing curve.[3]

In summary, we feel that these initial results are quite remarkable. They show 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. Potential Applications

So far, we have conducted this project as a purely scientific, basic research enterprise (as indeed we have the good luck of being able to afford, given the generous funding scheme). However, in the future, possibilities of practical application and exploitation of the results will also be looked at. After all, we are developing algorithms and computer programs that solve interesting musical problems, problems of possibly practical import, such as beat tracking, score extraction, and performance visualization.

For instance, *beat and tempo tracking* abilities are required in interactive music systems, be it

[3]To be more precise: the rules predict whether a note should be lengthened or shortened; the *precise numeric factor* of lengthening/shortening is predicted by a *k-nearest-neighbor algorithm* (with $k = 3$) that uses only instances for prediction that are covered by the matching rule, as proposed in [44,45].

in interactive performance on stage (e.g., for automatic synchronization with performers, lights, recording equipment, computer animations, video, etc.) or in systems that provide automatic accompaniment to soloists for practicing purpose (such as *Band-in-a-Box*). They are also important for audio content analysis systems, for example for the indexing and content-based retrieval of audio data in multimedia databases, or the automatic transcription of music. The problems of *quantization* as well as *pitch spelling* arise (and are not usually dealt with in a really intelligent way) in sequencers and music typesetting software.

We also plan to develop a *visualization system* for expressive performance and possibly, based on that, an *interactive performance editor*. The motivation for this is to facilitate efficient experimentation with different performance models in the context of our project, but again, there may be a tremendous application potential (e.g., in conservatories or in recording studios).

Whether the main product of the project — the *predictive models of expressive performance* — will have any immediate practical use remains to be seen. That is not the main motivation for this research, but we are open to interesting application ideas.

7. Preview of Planned Research

Expressive performance is an extremely complex phenomenon, and the above-mentioned research has just begun to scratch the surface of some of the problems. Many things remain to be done. To arrive at truly reliable results, substantial and systematic experiments must be performed with different expression dimensions, different representations of the example data (i.e., different music-structural vocabularies), different training corpora, different performers, and different learning algorithms.

A goal of central importance is the investigation of expressive patterns and strategies at *different structural levels* of musical organisation. That involves first identifying and separating expressive phenomena at different structural levels (e.g., at the level of notes, local segments, phrases), which is certainly a non-trivial problem, especially if one does not know the appropriate model classes for the different levels. The second step is then to induce individual models that explain as much as possible of the phenomena observed at each level. Finally, a strategy is needed for combining the predictions of the various models into overall expression profiles.

Another interesting question concerns the *interactions* between different expression dimensions (e.g., how does timing affect dynamics and articulation, and how are these dimensions used in conjunction by performers to create certain effects?). Again, this is a difficult problem and will require elaborate experimental designs.

Studies concerning commonalities and differences between *different musical styles* should be performed: which performance principles seem near-universal (at least with respect to 'classical' tonal music), which ones seem specific to musics of a particular style or period? If we manage to collect enough high-quality performance data from diverse styles, an empirical study in this direction should be feasible.

And finally, it would be fascinating to study *great artists*, on the basis of *sound recordings*. If we manage to refine our audio analysis algorithms to the point where they produce relatively precise timing measurements at detailed metrical levels, we will be able to study the timing in arbitrary sound recordings. We could then try to use learning algorithms to discover characterizations of (aspects of) the personal performance styles of great artists, and of commonalities and differences between different artists (along the lines of, e.g., [32], but with much larger and more diverse sets of examples, and with inductive learning algorithms that actually search for comprehensible descriptions of stylistic patterns).

We hope to be able to report on progress along some of these lines in the next progress report, some two years from now.

Acknowledgements

The project is made possible by a very generous START Research Prize (project no. Y99-INF) by the Austrian Federal Government, administered by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung (FWF)*. The cooperative research on staccato with Roberto Bresin of KTH, Stockholm, was supported by the EU project HPRN-CT-2000-00115 MOSART. The

Austrian Research Institute for Artificial Intelligence acknowledges basic financial support by the Austrian Federal Ministry for Education, Science, and Culture. I would like to thank my colleagues Emilios Cambouropoulos, Simon Dixon, and Werner Goebl for their cooperation and many fruitful and enjoyable discussions. We are particularly grateful to the pianists Roland Batik and Philippe Entremont for allowing us to use their performances for our investigations. Many thanks also to the L. Bösendorfer Company, Vienna, and in particular Fritz Lachnit for generous help with the Bösendorfer SE290.

References

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

[2] 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.

[3] 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.

[4] Cambouropoulos, E. (2001). Automatic Pitch Spelling: From Numbers to Sharps and Flats. In *Proceedings of the 8th Brazilian Symposium on Computer Music*, Fortaleza, Brazil.

[5] Cambouropoulos, E. (2001). Melodic Cue Abstraction, Similarity, and Category Formation: A Formal Model. *Music Perception*, 18(3) (in press).

[6] Cambouropoulos, E. (2001). *The Local Boundary Detection Model (LBDM) and its Application in the Study of Expressive Timing.* Submitted.

[7] Cambouropoulos, E., Smaill, A., and Widmer, G. (1999). A Clustering Algorithm for Melodic Analysis. In *Proceedings of the Diderot Forum on Mathematics and Music*, Vienna, Austria. Austrian Computer Society Press.

[8] Cambouropoulos, E. and Widmer, G. (2000). Automatic Motivic Analysis via Melodic Clustering. *Journal of New Music Research*, 29(4) (in press).

[9] Clarke, E. (1985). Some Aspects of Rhythm and Expression in Performances of Erik Satie's 'Gnossienne No. 5', *Music Perception* 2, pp.299–328.

[10] Cohen, W. (1995). Fast Effective Rule Induction. In *Proceedings of the 12th International Conference on Machine Learning*, Tahoe City, CA. San Francisco, CA: Morgan Kaufmann.

[11] Darden, L. (1997). Recent Work in Computational Scientific Discovery. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum.

[12] Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. In J. Kittler and F. Roli (Ed.), *First International Workshop on Multiple Classifier Systems* (pp. 1-15). New York: Springer Verlag.

[13] Dixon, S. (2000). A Lightweight Multi-Agent Musical Beat Tracking System. In *Proceedings of the 6th Pacific Rim International Conference on Artificial Intelligence (PRICAI'2000)*, Melbourne, Australia.

[14] Dixon, S. (2000). Extraction of Musical Performance Parameters from Audio Data. In *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia (PCM 2000)*, Sydney, Australia.

[15] Dixon, S. (2001). Automatic Extraction of Tempo and Beat from Expressive Performances. *Journal of New Music Research* (to appear).

[16] Dixon, S. (2001). *An Interactive Beat Tracking and Visualisation System.* Submitted.

[17] Dixon, S. and Cambouropoulos, E. (2000). Beat Tracking with Musical Knowledge. In *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI-2000)*, Berlin. IOS Press, Amsterdam.

[18] Friberg, A. (1995). *A Quantitative Rule System for Musical Performance.* Ph.D. dissertation, Department of Speech Communication and Music Acoustics, Royal Institute of Technology (KTH), Stockholm.

[19] Fürnkranz, J. (1999). Separate-and-Conquer Rule Learning. *Artificial Intelligence Review* 13(1).

[20] Goebl, W. (1999). Analysis of Piano Performance: Towards a Common Performance Standard? (Abstract). In *Proceedings of the 5th International Conference on Music Perception and Cognition (SMPC'99)*, Evanston, Illinois.

[21] Goebl, W. (2001). Melody Lead in Piano Performance: Expressive Device or Artifact? *Journal of the Acoustical Society of America* (in press).

[22] Goebl, W. and Dixon, S. (2001). Analysis of Tempo Classes in Performances of Mozart Sonatas. In *Proceedings of the 7th International Symposium on Systematic and Comparative Musicology*, Jyväskylä, Finland.

[23] Hartigan, J. (1975). *Clustering Algorithms.* Chichester, UK: John Wiley & Sons.

[24] Hunter, L. (ed.) (1993). *Artificial Intelligence and Molecular Biology.* Menlo Park, CA: AAAI Press.

[25] King, R.D., Muggleton, S., Lewis, R.A., and Sternberg, M.J.E. (1992). Drug Design by Machine Learning: The Use of Inductive Logic Programming to Model the Structure-activity Relationship of Trimethoprim Analogues Binding to Dihydrofolate Reductase. In *Proceedings of the National Academy of Sciences*, Vol. 89, pp.11322–11326.

[26] Kroiss, W. (2000). *Parameteroptimierung für ein Modell des musikalischen Ausdrucks mittels Genetischer Algorithmen.* Master's thesis, Department of Medical Cybernetics and Artificial Intelligence, University of Vienna.

[27] Muggleton, S., King, R.D., and Sternberg, M.J.E. (1992). Protein Secondary Structure Prediction Using Logic-based Machine Learning. *Protein Engineering* 5(7), pp.647–657.

[28] Narmour, E. (1992). *The Analysis and Cognition of Melodic Complexity: The Implication-Realization Model.* Chicago, IL: University of Chicago Press.

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

[30] Palmer, C. (1996). On the Assignment of Structure in Music Performance. *Music Perception* 14, 23–56.

[31] Quinlan, J.R. (1990). Learning Logical Definitions from Relations. *Machine Learning* 5, 239–266.

[32] 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.

[33] Repp, B. H. (1997). The Aesthetic Quality of a Quantitatively Average Music Performance: Two Preliminary Experiments. *Music Perception* 14, 419–444.

[34] 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.

[35] 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.

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

[37] 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.

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

[39] Shavlik, J.W., Towell, G., and Noordewier, M. (1992). Using Neural Networks to Refine Biological Knowledge. *International Journal of Genome Research* 1(1), pp.81–107.

[40] 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.

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

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

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

[44] Weiss, S. and Indurkhya, N. (1995). Rule-based Machine Learning Methods for Functional Prediction. *Journal of Artificial Intelligence Research* 3, 383–403.

[45] Widmer, G. (1993). Combining Knowledge-based and Instance-based Learning to Exploit Qualitative Knowledge. *Informatica* 17, 371–385. Special Issue on Multistrategy Learning.

[46] Widmer, G. (1995). Modelling the Rational Basis of Musical Expression. *Computer Music Journal* 19(2), pp.76–96, MIT Press.

[47] Widmer, G. (1996). What Is It That Makes It a Horowitz? Empirical Musicology via Machine Learning. In *Proceedings of the 12th European Conference on Artificial Intelligence (ECAI-96)*, Budapest. Wiley & Sons, Chichester, UK.

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

[49] Widmer, G. (1998). Applications of Machine Learning to Music Research: Empirical Investigations into the Phenomenon of Musical Expression. In R.S. Michalski, I. Bratko and M. Kubat (eds.), *Machine Learning, Data Mining and Knowledge Discovery: Methods and Applications.* Chichester, UK: Wiley & Sons.

[50] 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.

[51] Widmer, G. (2001). *Machine Discoveries: Some Simple, Robust Local Expression Principles.* Submitted.