Statistical Pattern Recognition for Prediction of Solo Piano Performance

Chris Chafe
Center for Computer Research in Music and Acourstics
Music Department, Stanford University
cc@ccrma.stanford.edu

Abstract

The paper describes recent work in modeling human aspects of musical performance. Like speech, the exquisite precision of trained performance and mastery of an instrument does not lead to an exactly repeatable performed musical surface with respect to note timings and other parameters. The goal is to achieve sufficient modeling capabilities to predict some aspects of expressive performance of a score.

1 Introduction

The present approach attempts to capture the variety of ways a particular passage might be played by a single individual, so that a predicted performance can be defined from within a closed sphere of possibilities characteristic of that individual. Ultimately, artificial realizations might be produced by chaining together different combinations at the level of the musical phrase, or guiding in real time a synthetic or predicted performance.

A pianist was asked to make recordings (in Yamaha Disklavier MIDI data format) from a progression of rehearsals during preparation of Charles Ives' First Piano Sonata for a concert performance. The samples include repetitions of an excerpt from the same day as well as recordings over a period of months. Timing and key velocity data were analyzed using classical statistical feature comparison methods tuned to distinguish a variety of realizations. Chunks of data representing musical phrases were segmented from the recordings and form the basis of comparison.

Presently under study is a simulation system stocked with a comprehensive set of distinct musical interpretations which permits the model to create artificial performances. It is possible that such a system could eventually be guided in real time by a pianist's playing, such that the system is predicting ahead of an unfolding performance. Possible applications would include performance situations in which appreciable electronic delay (on the order of 100's of msec.) is musically problematic.

Caroline Palmer's comprehensive review of studies of expressive performance [1] presents several points that bear importance for the present work. Foremost, she warns against “drawing structural conclusions based on performance data averaged or normalized across tempi.”

Several reports are mentioned in conjuntion with the exploration of structure-expression relationships. Of significance here is corroboration for the salience of phrase-level units in analyzing structures underlying performance. For example, errors in complex sequences when analyzed suggest that phrase structures influence mental partitioning. Errors tend not to interact across phrase boundaries. Also, phrases appear to be tied to their global context in different ways. As the present data show, some phrases appear to be "tempo invariant" where others scale according to tempo-based ratios.

Palmer states, “Each performer has intentions to convey; the communcative content in music performance includes the performers’ conceptual interpretation of the musical composition.” Accordingly, expressive variations are intentional and have been shown to possess a high degree of repeatibiliy in patterns of timing and dynamics. Performers are deliberate in applying devices to portray concepts such as louder dynamics used to strengthen unexpected structural or melodic events. Similarly, events with higher tension (in a tension / relaxation scheme) might be brought out by being played longer.

2 Data Collection and Preparation

Pianist George Barth, a Professor of Performance in the Stanford University Music Department provided the recordings. The Ives piece seems particularly versatile for present purposes because its interpretations are have little influence from any common stylistic practice. Barth prepared the performance over the course of four months with nearly daily practice. The first of five samples were collected over several weeks once he felt confident that he knew the notes.

An extract of the fifth movement shown in Figure 1 was targeted for study after an initial look at the data confirmed good stability across the five samples. The 55 note passage was performed flawlessly in each take and provided sufficient length and variation for purposes of the analysis. The pianist was unaware of the the choice of the extract, so as far as he was concerned he was recording a much longer excerpt of the movement, thus avoiding any likelihood of study-influenced effect on the performance.

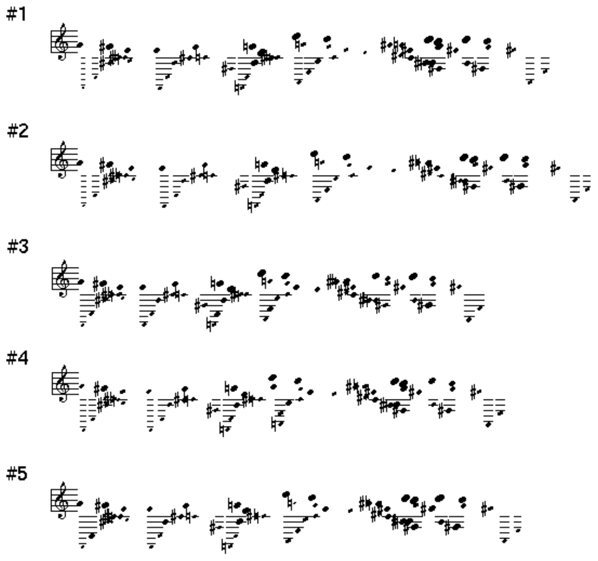

Figure 2: Displayed proportionally, the raw data for note onsets and key velocity shows expressive variations.

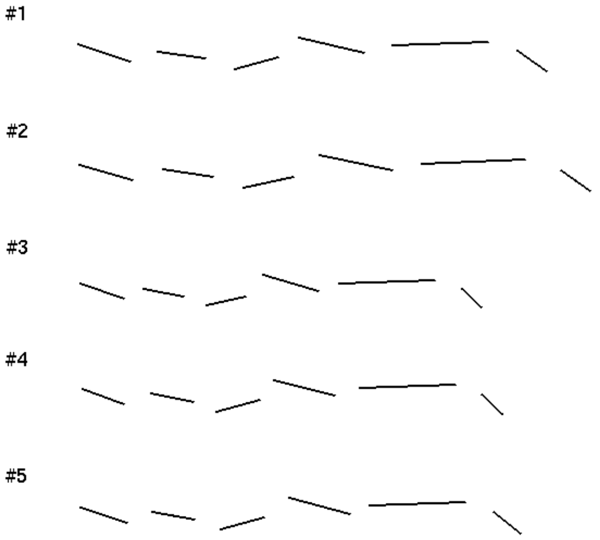

Figure 3: Sketching only phrase boundaries, tempo changes are visible both globally across phrases and internally within phrases.

Several steps were necessary to put the extract into a suitable form for analysis. The performances were recorded directly to the Disklavier's floppy disk in Yamaha's E-Seq MIDI data format. Conversion to Standard MIDI File Format type 1 was accomplished in software with Giebler Enterprises' DOMSMF utility. Segmentation of the extract and conversion to type 0 format utilized Opcode Systems' Vision sequencer. Trimmed and converted files were then imported into the Common Music Lisp environment for the first stages of analysis.

The present study is limited to note onset timings and key velocity (dynamic) information. Duration and pedaling data have been preserved during the conversion process for possible subsequent use.

Figure 2 is a proportional graph depicting the raw quantities recorded from the five perfomances. In Figure 3, phrase timing differences are highlighted by connecting a line segment between the positions of the starting and ending note-heads of each phrase.

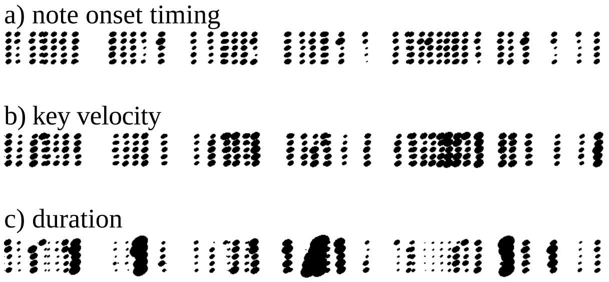

Figure 4: Variation in three parameters across the five performances.

For ease of comparison, Figure 4 isolates parameters with phrases aligned (by lining up events on the timings of the first performance and varying the notehead size according to the parameter). In b), variations of note onset timing use data relative to the first performance (larger noteheads indicate greater lengthening). Dynamic information is depicted by notehead sizes that depend on the key velocities found in each performance. Durational information is shown for informational purposes but was not analyzed further.

3 Features for Covariance Analysis

Performance data, being sequential, requires the choice of a time window relevant to the features that the analysis intends to capture. As can be seen in the above graphs of the raw data, phrase-level comparisons are of interest. Phrases have different overall durations and begin at different times according to the tempo of the performance. The first step in preparing features for classification was to isolate the phrases, setting the elapsed time of each event to be relative to the onset of the phrase rather than its absolute time.

The chosen feature dimensions of note onset timings and dynamics are expressed as differences from a reference performance. A less effective approach would be to express differences relative to perfect values derived from proportions in the score, which itself is a sort of performerless performance. Differences obtained against the score are distributed more coarsely; timings are relative to a less realistic baseline and the values for dynamics have to be guessed at (since they are specified only generally). By referencing to a recorded performance, differences distribute more usefully. Stylistic or habitual features such as phrase-final lengthenings are made implicit and dynamic differences are relative to actual values.

To compare two performances, three performances are required: the reference (P_{ref}) and the two inputs (P_1 and P_2). For each phrase, each event in each input is mapped according to the two feature dimensions. The intended result is that the inputs will be sufficiently distinguishable in this space. Figure 5 shows the distribution that results for the first three phrases from P_{ref} performance #5, P_1 #2, and P_2#3. A separator has been calculated based on the Mahalanobis distance to the center of each performance cluster. The separator as shown correctly classifies xxx% of the displayed points. For P-ref =#5, the four other performances are cross-compared and with a lowest statistic of xxx% correct for P-1 =#xxx, P2 =#xxx. The average statistic is xxx%.

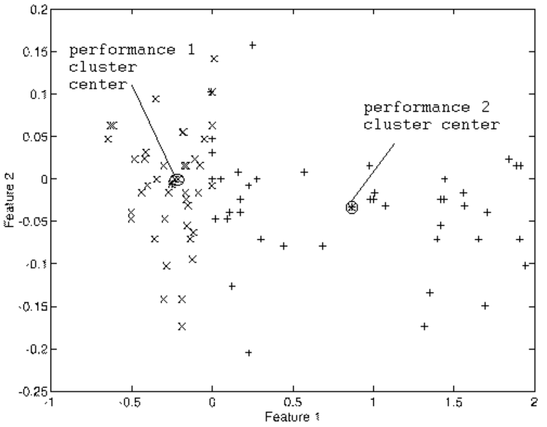

Figure 5: Note onset (feature 1) is plotted against key velocity (feature 2) for the first three phrases of two performances.

4 Other Clustering Methods

Nearest neighbor and K-means clustering algorithms permit... something to try.

5 Discussion

Covariance analysis can provide characterization of different performances within a comparison space. Phrase-by-phrase rhythmic and dynamic articulations can be successfully classified by applying various clustering algorithms. Performances that are not distinguishable by covariance analysis are presumed similar for the sake of the model being developed.

A future goal is to produce imitative expressive performances via behavior-based manipulation. A given phrase would be realized by selecting a stored phrase from the analyzed set of phrases. In a purely guided mode, the "operator" would determine the sequence of phrase samples, perhaps also choosing from interpolated combinations as in [2]. Another application involves real-time analysis / synthesis of expressive performance. A pianist performing in real time would be located in the above comparison space and on-the-fly classification decisions would predict the most likely stored performance matching the current input. The ability to predict ahead of a current performance can be useful, for example to overcome transmission delays.

The predict ahead capability of such a system is analgous to teleautonomous control in robotic applications [3]. In this sense, the remote instrument (robot) is played by its resident predictor (a remote simulator) guided by higher level controls transmitted to it by analysis of the local performer (human operator). To be agonizingly complete in this analogy, a remote accompanist's performance (environmental feedback) is provided to the local performer via a second identical system running back the other direction (by predicting a remote performer locally). In other words, a bi-directional setup might allow a piano duo to perform together across oceans. The two simultaneous concerts would differ, but not by much, assuming the analyzers and predictors are effective.

7 Conclusions

A performance is made of many layers. Global tempo changes and goals over longer structures remain to be described. The force-feedback manipulation of a performance described in the accompanying article [4] operates on the phrase-level substrate which has been the target of the present analysis. O’Modhrain’s control system displays the possible realizations of a given phrase within its comparison space. As a performance unfolds, the manipulator is guided through a dynamically changing scene.

Arkin describes layers of schema that operate in combination to enable guided teleautonomous behavior of a robot. “quote xxxxxxxxx” The same problem obtains here. By patterning phrase-level behavior according to a predictor, partially autonomous performance is possible which can be realized in conjunction with global schema. Control of these other layers a subject for future work, either in testing a real-time remote performance venue or addressing such issues in an editing environment for algorithmic performance.

8 Acknowledgments

Many thanks to George Barth for discussions and another round of experiments dissecting his excellent piano playing.

References

[1] Palmer, C. 1997. “Music Performance,” Ann. Rev. Psychol., 48, pp. 115-38.

[2] Chafe, C., S. O’Modhrain 1996. “xxx,” Proc. ICMC, Hong Kong, pp. xxx.

[3] O’Modhrain, S. 1997. “THE FUZZY MOOSE: A Haptic Tool for Tracking the performance of Fuzzy Classifiers in real-time.,” Proc. ICMC, Thessaloniki.

[4] Arkin, R.