Statistical Pattern Recognition for Prediction of Solo Piano Performance

Chris Chafe
Center for Computer Research in Music and Acoustics
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.” Data in the present work is analyzed in a way that preserves nuances until the final steps of classification.

Several reports are mentioned in conjunction with the exploration of structure-expression relationships and corroborate the salience of phrase-level units in performance analysis. 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. Some phrases appear to be "tempo invariant" where others scale according to tempo-based ratios.

In Palmer’s words, “Each performer has intentions to convey; the communicative content in music performance includes the performers’ conceptual interpretation of the musical composition.” Expressive variations are intentional and show a high degree of repeatability in patterns of timing and dynamics. Performers are deliberate in applying devices to portray their concepts, for example choosing louder dynamics to strengthen unexpected structural or melodic events. Events with higher tension (in a tension / relaxation scheme) might be brought out by being played longer.

2 Data from Rehearsals

Pianist George Barth, a Professor of Performance in the Stanford University Music Department, provided the recordings. He prepared his performance over the course

of four months with nearly daily practice. The first five samples that are analyzed here were collected over several weeks, beginning after he felt confident of the notes.

An extract of the fifth movement 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 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.

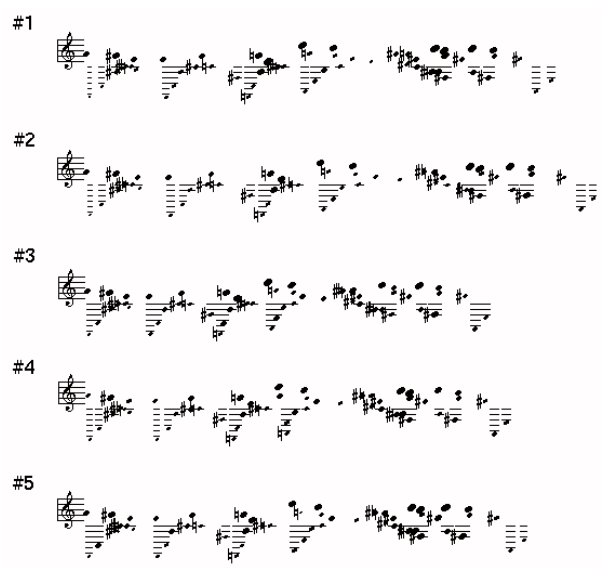

Figure 1: Displayed proportionally, the raw data for note onsets and key velocity shows expressive variations.

Several steps were necessary to prepare the extract 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 was accomplished with 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 1 shows proportionally the raw quantities recorded from the five performances. In Figure 2, phrase timing differences are depicted by connecting a line segment between the positions of the starting and ending note-heads of each phrase.

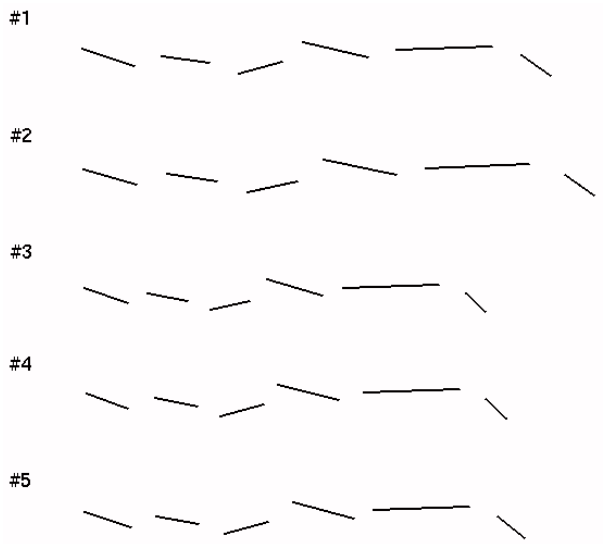

Figure 2: Sketching only phrase boundaries, tempo changes are visible both globally across phrases and internally within phrases.

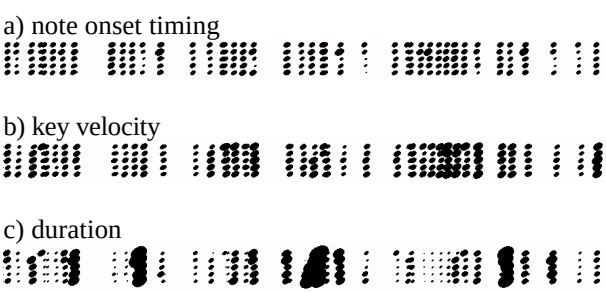

Figure 3: Variation in three parameters across the five performances.

For ease of comparison, Figure 3 isolates parameters with phrases aligned (by lining up events on the timings of the first performance and varying the notehead size according to the given 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. Note durations are shown for informational purposes but were not analyzed further.

3 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. Because phrases have different overall durations and begin times which are influenced by 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 two features chosen as dimensions for a covariance analysis are note onset timings and dynamics expressed as differences from a reference performance (key velocities are scaled to a range of 0 - 1). 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 values for dynamics have to be intuited (since they are specified only generally). By referencing to a recorded performance, differences are distributed 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 4 shows the distribution that results for the fifth phrase with P_{ref} as performance #5, P_1 as #1, and P_2 as #2. A separator has been calculated based on the Mahalanobis distance to the center of each performance cluster [2]. The separator as shown correctly classifies 76% of the displayed points.

As the performance unfolds, the relative positions of cluster centers change phrase-by-phrase. Figure 5 shows trajectories mapped for four performances during the second half of the excerpt.

The analysis demonstrates an ability to identify nearby performances. In Figure 6, a coincidentally close pair of performances of one phrase was correctly classified.

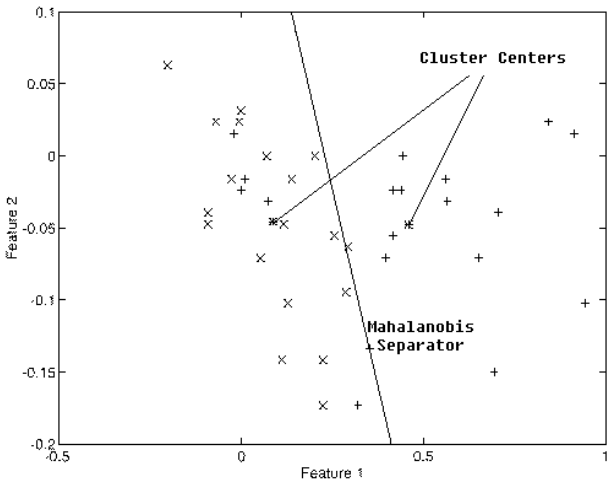

Figure 4: Note onset timing (feature 1) is plotted against key velocity (feature 2) for the same phrase in two performances. Quantities are differences from values for the same notes in a third, reference performance.

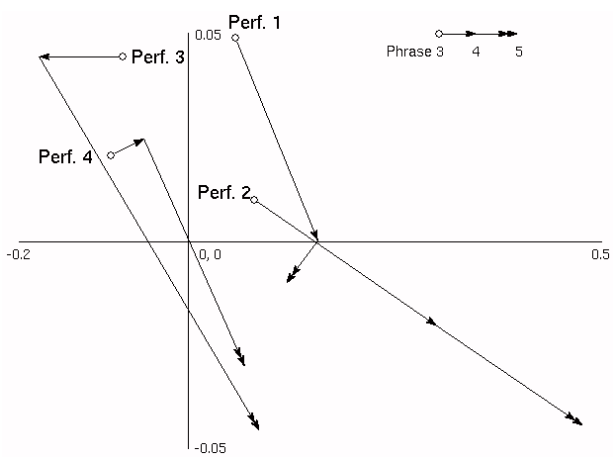

Figure 5: The relative positions of cluster centers change phrase-by-phrase. The trajectories of four performances are shown for three phrases in the same feature comparison space as Figure 4.

4 Discussion

Phrase-by-phrase tendencies in rhythmic and dynamic articulations can be successfully classified by covariance analysis. Performances that are not distinguishable are presumed similar for the sake of the model being developed.

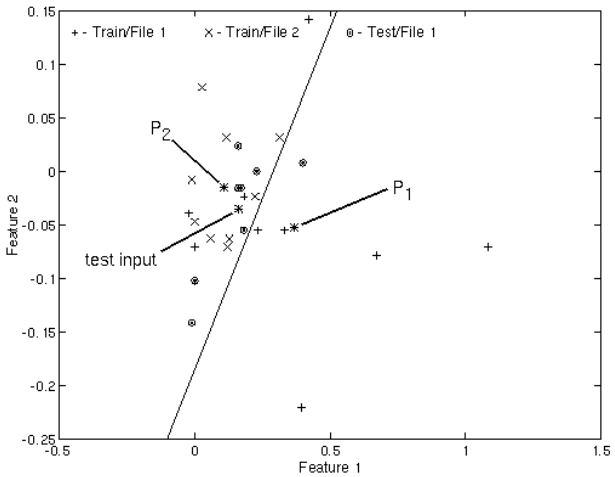

Figure 6: Successful classification of an "unknown" performance of phrase 4 in a comparison space created with P_{ref} = #5, P_1 = #1, and P_2 = #3.

A future interest is to produce imitative expressive performances via behavior-based manipulation. A given passage would be realized by selecting a stored phrase from an 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 [3]. Another mode involves real-time analysis / synthesis of expressive performance. A pianist performing in real time would be located in the 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 is analogous to teleautonomous control in robotics applications [4]. The remote instrument (robot) is played by its predictor (a remote simulator) guided by 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 back to the local performer via a second system running in the other direction. 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.

Force-feedback manipulation of the model is discussed in O'Modhrain's accompanying article [5]. Her system operates on the phrase-level substrate that has been the focus of the present analysis and is intended to display the possible realizations of a given phrase within its comparison space. As a performance unfolds, the manipulator is guided through a dynamically changing scene, much like Figure 5.

A performance is made of many layers. Global tempo changes and other longer structures remain to be described in the present model. Arkin describes layers of schema operating in combination to enable guided teleautonomous behavior of a robot. "…that schema-based reactive control results in a 'sea' of forces acting upon the robot." By patterning phrase-level behavior according to a predictor, partially autonomous performance is possible which can be realized in conjunction with global and other performance schema. Control of these other layers is a subject for future work, either in testing a real-time remote performance venue or in an editing environment for algorithmic performance.

5 Acknowledgments

Many thanks to George Barth for discussions and another round of experiments dissecting his excellent piano playing. Also to participants of the HCI Design Course Fall '96 at San Jose State, Stanford and Princeton Universities.

References

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

[2] Devroye, L., et al. 1996. *A Probabilistic Theory of Pattern Recognition*, New York: Springer-Verlag.

[3] Chafe, C., S. O'Modhrain 1996. "Musical Muscle Memory and the Haptic Display of Performance Nuance," Proc. ICMC, Hong Kong, pp. 428-31.

[4] Arkin, R. 1991, "Reactive Control as a Substrate for Telerobotic Systems," IEEE AES Sys. Mag., pp. 24-31.

[5] O'Modhrain, S. 1997. "Feel the Music: Narration in Touch and Sound," Proc. ICMC, Thessaloniki.