CENTER FOR COMPUTER RESEARCH IN MUSIC AND ACOUSTICS
AUGUST 1997

DEPARTMENT OF MUSIC
REPORT NO. STAN-M-101

CCRMA PAPERS PRESENTED AT THE 1997 INTERNATIONAL COMPUTER MUSIC CONFERENCE THESSALONIKI, GREECE

Jonathan Berger, Chris Chafe, Alex Igoudin, David Jaffe, Tobias Kunze, Scott Levine, Fernando Lopez-Lezcano, Sile O'Modhrain, Pat Scandalis, Gary Scavone, Julius Smith, Tim Stilson, Heinrich Taube, Scott Van Duyne, Tony Verma

CCRMA
DEPARTMENT OF MUSIC
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-8180

TABLE OF CONTENTS

A Neural Network Model of Metric Perception and Cognition in the Audition of Functional Tonal Music.

Jonathan Berger
CCRMA
Stanford University
Stanford, CA 94305, U.S.A
brg@ccrma.stanford.edu

Dan Gang *
Institute of Computer Science
Hebrew University
Jerusalem 91904, Israel
dang@cs.huji.ac.il

Abstract

In our previous work we proposed a theory of cognition of tonal music based on control of expectations and created a model to test the theory using a hierarchical sequential neural network. The net learns metered and rhythmecized functional tonal harmonic progressions allowing us to measure fluctuations in the degree of realized expectation (DRE). Preliminary results demonstrated the necessity of including metric information in the model in order to obtain more realistic results for the model of the DRE. This was achieved by adding two units representing periodic index of meter to the input layer. In this paper we describe significant extensions to the architecture. Specifically, our goal was to represent more general meter tracking strategies and consider their implications as cognitive models. The output layer of the sub-net for metric information is fully connected to the hidden layer of sequential net. This output layer includes pools of three and four units representing duple and triple metric indices. Thus the sub-net was able to influence the resulting DRE, that was expected by the net. Moreover, by including multiple metric parsings in the output layer the net reflects conflicts between parallel possible interpretations of meter. This output was fed back into the sub-net to influence the next predictions of the DRE and the meter. In addition, the target harmony element was fed into the context instead of the actual output, thus simulating the interactive influences of harmonic rhythm and meter.

1 Introduction

"The poets have a proverb: Metra parant animos (the emotions are animated through verse). They say such quite rightly: for nothing penetrates the heart as much as a well-arranged rhyme scheme [Mat39]".

Johann Mattheson's awareness of the cognitive power of underlying metric temporal patterns (be it musical metric feet or rhythmic modes) in music and poetry has been consistently stated and, over the past century, empirically researched. That listening to music involves an initial creation of a metric schema has been well documented. What is not clear, however, is the process in which the listener arrives at a working schema.

In this paper we explore and model a possible scenario of metric decision making. As a point of departure we incorporate observations, speculation, and perceptual studies that suggest:

1. Constructing a metric schema is a task critical to music audition. In Mattheson's words "...the ordering of the feet in poetry and the well-constructed alternation of meters, even if there were no rhyme scheme, produces something initially so certain and clear in the hearing that the mind enjoys a secret pleasure from the orderliness and accepts the performance so much the easier."

2. Listeners of Western music have preconceived organizational schemas grouping into duple or triple metric units. Listeners count in hierarchies (base 3 or base 4 for most common meters). [Pov81] demonstrated that untrained listeners can accurately distinguish between duple and triple metric units. Furthermore, considerable evidence of preconceived grouping preferences suggest that this is applicable to meter recognition.

*Dan Gang is supported by an Eshkol Fellowship of the Israel Ministry of Science

Although generative algorithms (e.g., [LHL82]) and autocorrelative methods (e.g., [DH89]) for meter recognition are successful in their task they do not offer a plausible explanation of how a listener applies schematic based expectation of duple or triple groupings to determine meter. The music theory literature regarding meter (e.g., [LJ83]) similarly fail to account for this basic task.

3. Metric awareness is necessary in building a network of implications and expectations which lies at the heart of the musical experience. London [Lon92] proposes that metric cognition involves a two stage process comprising a recognition phase (establishment of a metrical framework) and a continuation phase (projection of the chosen framework into the future). Thus meter is critical in establishing expectations. London maintains that most computational and experimental studies of meter regard the recognition stage while theoretical studies provide retrospective evidence. Implied here is a failure to provide an adequate study of metric recognition that incorporates prediction and continuation. Our experiments take this challenge as a point of departure.

We propose that a listener simultaneously activates two parallel metric schemas each with some degree of independence. When one proves to correlate more consistently with other incoming patterns (dynamic accentuation, harmonic accent, phrase and articulation accents, etc.) the metric schema that fails 'turns off'. Furthermore, our model enables the integration of mutual influences of two interrelated aspects of musical expectations: schematic metric awareness (which influences functional tonal expectations) and learned functional tonal implications that in and of themselves create metric expectations. The merger and integration of these cognitive processes allow for a more refined model of music audition.

2 The network design

2.1 Architecture of the network

In our previous model of fluctuation in DRE (see [GB96] and [BG96]), we adopted a three-layer sequential net in which 12 state units establish the context of the current chord sequence, and the 12 output layer units represent the prediction of the net for the subsequent chord. Both, the state and the output units are pitch class (PC) representations of triads and tetrads in the sequence. The output layer is fed

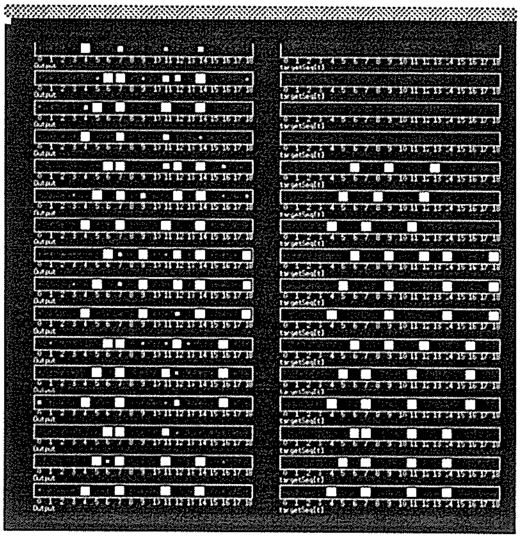


Figure 1: Simulation of expectancies. From left to right - 4 units represent duple meter and 3 more represent triple meter, the last 12 units are the harmonic expectations represented by 12 PCs. The size of the squares is proportional to the strength of the units' activity. Time proceeds from bottom-up. The right column represents the input and the left column visualizes the net's prediction for the meter and harmony. The progression is -
[3/4 I I I — vi vi ii — V V V7 — I I I]

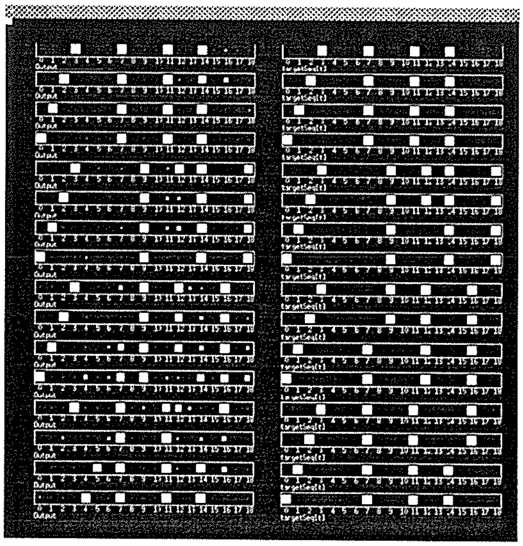


Figure 2: The progression is -
[4/4 I I vi vi —IV IV ii ii —V V V7 V7 —I I I I]

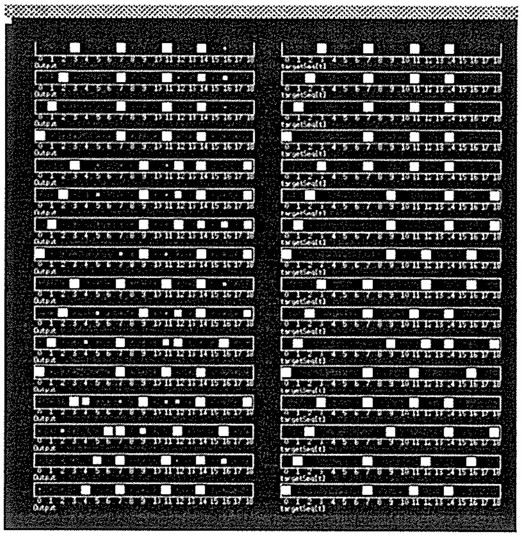


Figure 3: The progression is -
[4/4 I IV V I —vi V7 I IV —ii V V I —I I I I]

back into the state units to influence the next prediction of the net. The value of the state units at time t is the sum of its value at time $t-1$ multiplied by decay parameter and value of the output units at time $t-1$. By integrating a sub-net with the sequential net we supplemented the model with a simple metrical organizer that supplied a periodic beat stream of four beats per measure of duple harmonic progressions.

This model is extended by adding triple meter patterns to the architecture. In so doing we examine how metric expectations can influence the harmonic predictions and how the harmonic progression together with the context of the meter influence the prediction of meter.

This architecture differs from the previous model in a number of respects. The representation of meter is extended. We incorporate into the net's state units two pools (or a sub-net) of: 4 units to represent duple meter and 3 more to represent triple meter. These units are connected to the hidden layer together with the pool of the PCs representing the harmonic context. The hidden units are connected to the output layer. The output layer contains three pools of units: a pool for the harmonic expectations represented by the 12 PCs; and the two pools to represent expectations for duple and triple meter. The output of the 7 units of the meter is fed back into the corresponding pools of the state. The output of the prediction of the net for the harmonies' expectation were used to measure DRE and the target was fed into the PC units of the state, to establish the current harmonic context. We thus model the mutual influences of harmony on meter and meter on harmony. We note the enhancement of this method in quantifying the DRE. The DRE is also influenced by the metric expectations. This is particularly evident in (fig 4) where conflicting metric information greatly affected the DRE.

2.2 The set of learning examples

We use a learning set of functional tonal harmonic patterns. The patterns were evenly divided into duple and triple meter progressions. Harmonic rhythm in the learning set ranged from one chord per measure to one chord per beat, although the weighting was on one and two chord changes per measure for both duple and triple patterns.

3 Running the Net

3.1 The Learning Phase

For the learning phase the net was given thirty examples containing duple and triple patterns of harmonic progressions. After training, the net was able to reproduce the examples. We have tested the performance of the network with several different learning parameters. For example we found that for this task the net required relatively high value for the decay parameter.

3.2 The Generalization Phase

In this phase the net was given four new sequences. The target sequence was compared to the current harmonic and metric prediction of the net. The meter was fed back into the meter's pools of the state units and the target of the current harmonies was fed into the PC units. In analyzing the output we consider the distribution of the units' activation. By calculating how much of the target is present in the harmony pool of the output units, we were able to suggest a quantitative measurement of the DRE. The units of the meter pools in the output reflect duple and triple interpretations and clearly demonstrate conflicting metric and harmonic information.

4 Data Analysis

4.1 Figure 1:

[3/4 I I I — vi vi ii — V V V7 — I I I]

This example represents the output of a standard four measure progression in triple meter. The progression should show a high DRE. The role of the metric subnet is critical in the network's agility in detecting the correct harmonic rhythm by beat five. Of note is the openness of the system to change on beat three (resulting from the inconclusive assistance of the metric sub-net). However the downbeat of measure two entrains the network by supposing a metric schema which fully conditions expectation for harmonic progression and change. Thus, in measure two the expectation for a subdominant harmony is progressively strengthened and the expectation for a change to the dominant is highly expected. (The inconclusive expectation for tonic continuance in the final measure is an artifact of 'padding' the example in order to incorporate longer progressions).

4.2 Figure 2:

[4/4 I I vi vi —IV IV ii ii —V V V7 V7 —I I I I]

In this example a harmonic progression in 4/4 with a high DRE is input as a target sequence. In this example the initial willingness for change on beat three (evident in the distribution of strength of PC7 to PC5 and PC9 representing an expectation for shift to the sub dominant) is immediately followed in beat four by an even stronger expectation for change to a subdominant. The lowest DRE in the entire progression occurs in beat five. Here, the downbeat is fully recognized as a point of harmonic shift, with a greater expectation for sub dominant harmony, but with an openness for a dominant downbeat. The arrival of a subdominant in correspondence to the metric downbeat sets a strong expectation for the completion of the progression.

4.3 Figure 3:

[4/4 I IV V I —vi V7 I IV —ii V V I —I I I I]

In this example a distinct conflict between harmonic rhythm and meter results in significant drops in DRE. The hastened harmonic rhythm (a chord already on the second beat, setting up a quarter note harmonic rhythm) is resisted in the output's expectation for continued subdominant harmony in beat 3. The arrival of a tonic on beat four of measure one throws both the metric counter and the harmonic expectations into flux. The drop in DRE is particularly interesting in that the distribution of expectations is not willy nilly but rather reflective of an ambiguity, in which conflicting functional regions (tonic/ dominant) are confused. This conflict persists until the final measure.

5 Discussion

Some basic questions regarding the perception of meter in tonal music are raised. Specifically:

1. How does a listener identify the meter, when hearing an unfamiliar work?
2. Is the process of metric cognition one of parallel or sequential testing? That is, do we consider multiple possible meters simultaneously, or do we test one and, failing to achieve a good 'fit', shift to another metric count?
3. What are the implications of these questions on our theory of musical expectations?

In our first experiment we extended the initial model by incorporating two parallel and independent counters for three beats and four beats. An experiment currently being considered is to commence with two parallel counters but shut one off when a strong correlation between a high DRE and one of the two pools in the metric sub-net is established. A second experiment under current consideration involves a change of data structure, such that multiple metric possibilities are reflected within a single counter.

References

[BG96] J. Berger and D. Gang. Modeling musical expectations: A neural network model of dynamic changes of expectation in the audition of functional tonal music. In *Proceedings of the Fourth International Conference on Music Perception and Cognition*, Montreal, Canada, 1996.

[DH89] P. Desain and H. Henkjan. The quantization of musical time: A connectionist approach. *Computer Music Jouranal (CMJ)*, 13(3), 1989.

[GB96] D. Gang and J. Berger. Modeling the degree of realized expectation in functional tonal music: A study of perceptual and cognitive modeling using neural networks. In *Proceedings of the International Computer Music Conference*, Hong Kong, 1996.

[LHL82] H. C. Longuet-Higgns and C. S. Lee. The perception of musical rhythms. *Perception*, 11:115–128, 1982.

[LJ83] F. Lerdahl and R. Jackendoff. *A Generative Theory of Tonal Music.* Cambridge (MA): MIT Press, 1983.

[Lon92] J. London. The cognitive implications of a dynamic theory of meter. In *Proceedings of the Fourth International Conference on Music Perception and Cognition*, Pennsylvania, 1992.

[Mat39] J. Matheson. *Der Vollkommene Cappelmeister.* Hamburg: Christian Herold, 1739.

[Pov81] D. Povel. The internal representation of simple temporal patterns. *Journal of Experimental Psychology: Human Perception and Performance*, 7:3–18, 1981.

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."
Data in the present work is analyzed in a way that preserves nuances until the final steps of classification.

Several reports are mentioned in conjuntion 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.

Palmer states, "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 repeatibiliy 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 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.

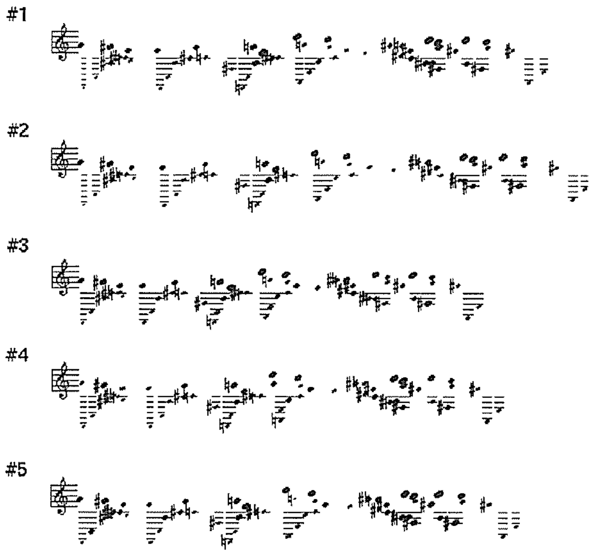


Figure 1: Displayed proportionally, the raw data for note onsets and key velocity shows expressive variations.

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

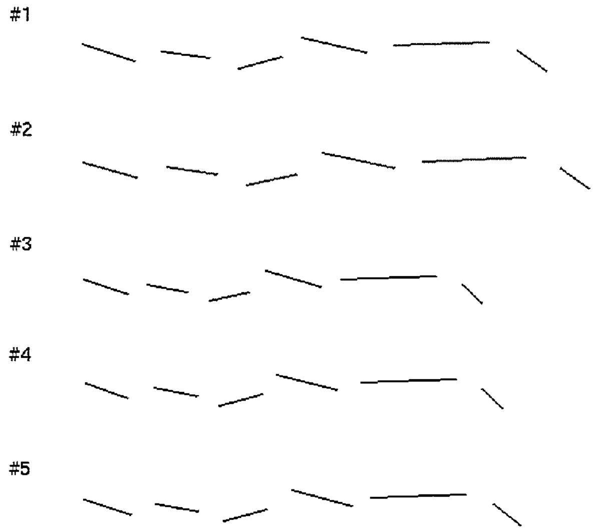
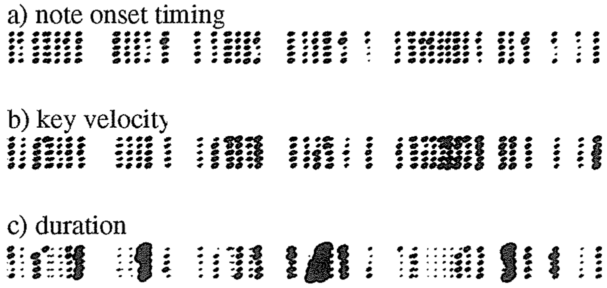


Figure 2: Sketching only phrase boundaries, tempo changes are visible both globally across phrases and internally within phrases.


a) note onset timing

b) key velocity

c) duration

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 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 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 times and 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 (P1 and P2). 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 Pref as performance #5, P1 as #1, and P2 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 correctly classify nearby performances. In Figure 6, a coincidentally close pair of performances for one phrase was correctly classified.

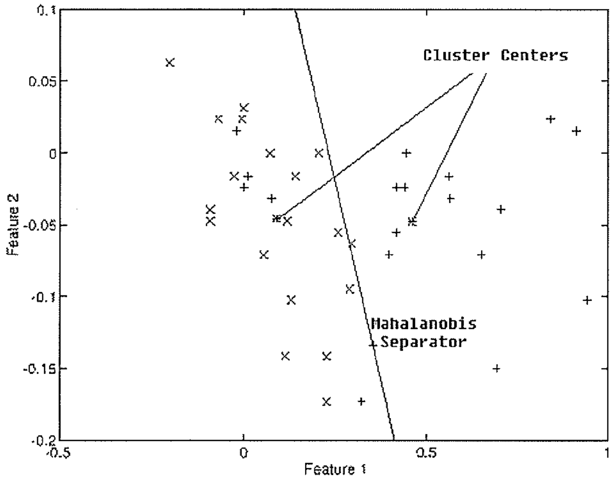


Figure 4: Note onset timing (feature 1) is plotted against key velocity (feature 2) for the same phrase in two performances. Quantities are differences from values for the same notes in a third, reference performance.

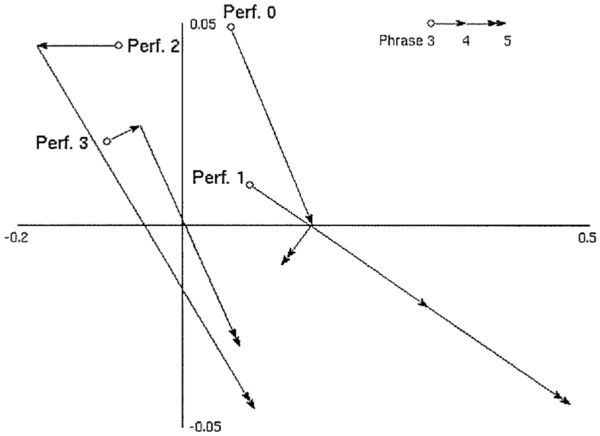


Figure 5: The relative positions of cluster centers change phrase-by-phrase. The trajectories of four performances are shown for three phrases in the same feature comparison space as Figure 4.

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.

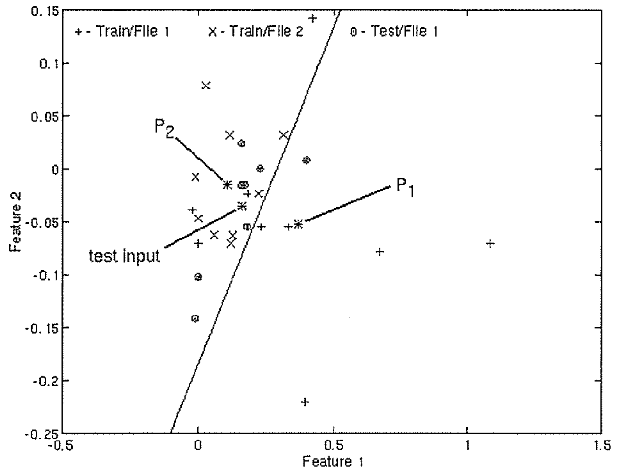


Figure 6: Successful classification of an "unknown" performance of phrase #4 in the comparison space of performances 1 and 3.

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 analgous to teleautonomous control in robotic 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.

A performance is made of many layers. Global tempo changes and other longer structures remain to be described in the present model. 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.

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

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

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