Modeling Moods in Violin Performances

Alfonso Perez, Rafael Ramirez, Stefan Kersten
Universitat Pompeu Fabra
Music Technology Group
Barcelona, Spain

***Abstract*—In this paper we present a method to model and compare expressivity for different Moods in violin performances. Models are based on analysis of audio and bowing control gestures of real performances and they predict expressive scores from non expressive ones.**

Audio and control data is captured by means of a violin pickup and a 3D motion tracking system and aligned with the performed score.

We make use of machine learning techniques in order to extract expressivity rules from score-performance deviations. The induced rules conform a generative model that can transform an inexpressive score into an expressive one.

The paper is structured as follows: First, the procedure of performance data acquisition is introduced, followed by the automatic performance-score alignment method. Then the process of model induction is described, and we conclude with an evaluation based on listening test by using a sample based concatenative synthesizer.

I. Introduction

Different approaches are found in the literature for modeling expressive performances: Fryden[4] tries an analysis-by-synthesis approach, consisting of a set of proposed expressive rules that are validated by synthesis. In [3] mathematical formulae is proposed to model certain expressive ornaments. Bresin[2] and Widmer[9] make use of machine learning in order to extract expressive patterns from musical performances. In [7] they use Case Based Reasoning, that is, a database of performances that conform the knowledge of the system. In this work we follow the work done by [8], also using machine learning techniques and more specifically inductive logic programming (ILP from now on), that has the advantage of automatically finding expressive patterns without the need of an expert in musical expressivity. Regarding research in generative models, in [5] a computational model of expression in music performance is proposed.

In general this techniques try to model perceptual features such as timing deviations, dynamics or pitch. In addition, we also inform the model with control gestures, more specifically bow direction and finger position.

Apart from calculating prediction errors, models are also evaluated by listening with the help of a sampled based concatenative synthesizer under development.

Four moods are analyzed: Sadness, Happiness, Fear and Anger. Expressive features analyzed are: tempo and a set note level descriptors: onset, note duration, energy, bow direction and string being played.

In the following sections we introduce the data acquisition procedure, we detail how the model is induced and how is it performing.

II. Data Acquisition and Analysis

The training data used in our experimental investigations consist of short melodies performed by a professional violinist in four Moods: Sadness, Happiness, Fear and Anger. Pieces were played twice with and without metronome.

A set of audio and control features is extracted from the recordings and stored in a structured format. The performances are then compared to their corresponding scores in order to automatically compute the performed transformations.

The main characteristic of our data acquisition system is that of providing also motion information. This information is used for learning the model as well as for the alignment and segmentation of the performances with the scores.

A. Scores

Scores are represented as a series of notes with onset, pitch (in semitones) and duration. No extra indications are given to the performer except for the Mood. They are used to calculate performance deviations from nominal attributes of the melody.

B. Audio acquisition

Audio is captured by means of a violin bridge pickup. This way we obtain a signal not influenced by the resonances of the violin acoustic box and the room, which makes segmentation much easier than if using a microphone. From the captured audio stream we extract the audio perceptual features: frame-by-frame energy, fundamental frequency estimation and aperiodicity function. Energy is used as input for learning the model.

C. Gesture acquisition and parameter calculation

Bowing motion data is acquired by means of two 3d-motion tracker sensors, one mounted on the violin and the other on the bow as we already described in [6]. We are able to estimate with great precision and accuracy the position of the strings, the bridge and the bow. With the collected data we compute, among others, the following bowing performance parameters: bow distance to the bridge, bow transversal position, velocity and acceleration, bow force and string being played. Bow direction change and playing string are used for the segmentation and as input for learning the model.

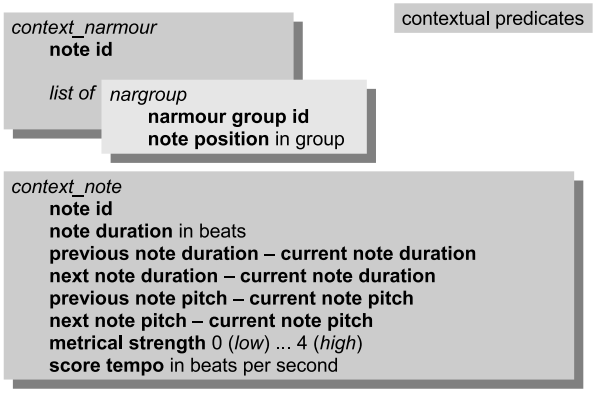

Fig. 1. Contextual predicates.

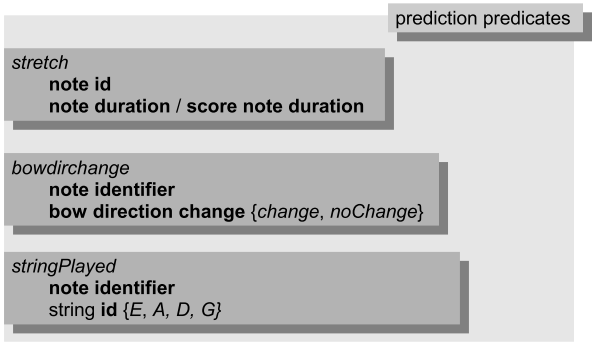

Fig. 2. Induction and Prediction predicates.

D. Score-Performance Alignment

Performances are represented with the same symbolic description as the score so that they can be aligned and deviations from the score obtained. An automatic alignment is carried out following [**?**]. It uses score information, bowing controls, and audio descriptors: A bow-direction change or a playing-string change indicates a note onset. In legato, notes segmentation is based on pitch and energy. Offsets are calculated by using energy levels. Automatic segmentation is finally manually corrected.

III. Expressive performance model

In this section we describe our inductive approach for learning the model by applying ILP techniques and we describe the evaluation results.

A. Data Description

After the alignment and segmentation, scores and expressive deviations of the performance are defined in a structured way using first order logic predicates. The musical context of each note is defined with the following predicates (Figure 1): *context_note* specifies information both about the note itself and the local context in which it appears. Information about intrinsic properties of the note includes note duration and note's metrical position, while information about its context includes the duration of previous and following notes, extension and direction of the intervals between the note and both the previous and the subsequent note, and tempo of the piece in which the note appears; *context_narmour* specifies the Narmour groups to which a particular note belongs, along with its position within a particular group. The temporal aspect of music is encoded via the predicates *pred* and *succ*. For instance, *succ(A,B,C,D)* indicates that note in position D in the excerpt indexed by the tuple (A,B) follows note C.

Expressive deviations in the performances are encoded using 4 predicates (Figure 2): *stretch* specifies the stretch factor of a given note with regard to its duration in the score; *bowdirchange* identifies points of change in bow direction; *stringPlayed* specifies in which string a note was played in the performance (certain pitches can be played in different strings resulting in a different timbre); and *dynamics* specifies the mean energy of a given note. These 4 predicates are also used for model prediction.

The use of first order logic for specifying the musical context of each note is much more convenient than using traditional attribute-value (propositional) representations. Encoding both the notion of successor notes and Narmour group membership would be cumbersome using a propositional representation. In order to mine the structured data we used Tilde's top-down decision tree induction algorithm ([1]). Tilde can be considered as a first order logic extension of the C4.5 decision tree algorithm: instead of testing attribute values at the nodes of the tree, Tilde tests logical predicates. This provides the advantages of both propositional decision trees (i.e. efficiency and pruning techniques) and the use of first order logic (i.e. increased expressiveness). The increased expressiveness of first order logic not only provides a more elegant and efficient specification of the musical context of a note, but it provides a more accurate predictive model.

B. Model Evaluation

We obtained correlation coefficients of 0.80 and 0.83 for the duration transformation and note onset prediction tasks respectively and we obtained a correctly classified instances percentage of 82% and 86% for the bow direction and string played prediction. These numbers were obtained by performing 10-fold cross-validation on the training data.

Additionally to the model performance error coefficients, listening tests are also carried as a perceptual evaluation of the models. For this we make use of a sample-based spectral concatenative synthesizer.

IV. Conclusions

We presented a model for expressive performances based not only on perceptual features but also informed with bowing. We introduced the procedure to acquire the data, learn the model and synthesize its predictions. The results seem to capture the expressive features performed. We obtained high prediction correlation coefficients and realistic synthesis of predicted performances.

V. Acknowledgments

This work was partly supported by the Spanish Ministry of Education and Science under Grant TIN2006-14932-

C02-01 (ProSeMus Project) and Yamaha Corp.

REFERENCES

[1] H. Blockeel, L. D. Raedt, and J. Ramon. Top-down induction of clustering trees. In *Proceedings of the 15th International Conference on Machine Learning*, 1998.

[2] R. Bresin. An artificial neural network model for analysis and synthesis of pianists performance styles. *JASA*, 105(2):1056, 1999.

[3] M. Clynes. *SuperConductor: The Global Music Interpretation and Performance Program*, 1998.

[4] L. Fryden, J. Sundberg, and A. Askenfelt. What tells you the player is musical? an analysis-by-synthesis study of music performance. *Publication issued by the Royal Swedish Academy of Music*, 39:61–75, 1983.

[5] P. N. Juslin, A. Friberg, and R. Bresin. Toward a computational model of expression in music performance: The germ model. *Musicae Scientiae*, Special Issue:63–122, 2002.

[6] E. Maestre, J. Bonada, M. Blaauw, A. Perez, and E. Guaus. Acquisition of violin instrumental gestures using a commercial emf device. In *Proceedings of International Computer Music Conference*, Copenhagen, Denmark, 2007.

[7] R. Mantaras, X. Serra, and J. L. Arcos. Saxex: A casebased reasoning system for generating expressive musical performances. In *Proceedings of International Computer Music Conference*, 1997.

[8] R. Ramirez, A. Hazan, E. Maestre, and X. Serra. A genetic rule-based expressive performance model for jazz saxophone. *Computer Music Journal*, 32(1):338–350, 2008.

[9] G. Widmer. Learning about musical expression via machine learning: A status report. In *17th National Conference on Artificial Intelligence*, 2000.