EXPRESSIVE IRISH FIDDLE PERFORMANCE MODEL INFORMED WITH BOWING

Alfonso Perez, Esteban Maestre, Stefan Kersten, Rafael Ramirez
Music Technology Group
Universitat Pompeu Fabra

ABSTRACT

We propose an expressive performance model for celtic fiddle based on the analysis of audio and bowing gestures of real performances. Existing expressivity models deal with perceptual features such as timing deviations or dynamics, but in some cases perceptual features are not enough. We propose a model not only informed with perceptual features but also with bowing gestures which are acquired by means of a 3D motion tracker system. Based on a set of expressive audio features and control gestures we apply machine learning techniques in order to induce an expressive performance model. We use the model to synthesize expressive celtic violin performances from inexpressive score descriptions. In this paper we describe the process of acquiring expressive performance features (both audio and gestures), we detail the automatic performance-score alignment and segmentation, we show how the model is induced, and finally we evaluate the results using a sample based concatenative synthesizer.

1. INTRODUCTION

Modeling expressive performances is an active research topic. Fryden[5] 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[11] make use of machine learning in order to extract expressive patterns from musical performances. In [8] 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 [10], 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.

In general this techniques try to model expressive features such as timing deviations, dynamics or pitch in a perceptual domain, that is, they try to model how we listen an expressive performance. Here we propose to inform the model not only with perceptual features but also with bowing gestures controlling the violin, that is, try to model what is the violinist doing in order to perform expressively.

It is also important to give a mean of evaluating the model. Predicted features of the model feed a sample based concatenative synthesizer so that listening tests can be carried out.

In this work we focus on the analysis on Irish Jigs and we concentrate on modeling note-level deviations, namely timing and energy as well as bowing (bow direction changes) and two expressive ornaments that appear often in Irish fiddling: *mordents* and *bowed triplets*.

In the following sections we introduce Irish fiddling, we describe how audio and gestural data is acquired and aligned in order to find performance deviations from score, we present the model, indicating how is it induced and what is predicting, we explain the synthesis stage and we conclude by giving some guidelines for future work.

2. FIDDLING IN IRISH MUSIC

Irish music comprises lots of different styles (Donegal, Sligo, etc.), musical forms (Reels, Jigs, Hornpipes, etc.) and expressive ornaments that can be performed by fingering (cut notes, rolls, hammer-ons, slides, etc.) or by bowing (slurs, dynamics, accents doublestops, shuffle patterns, bowed triplets, etc.). Most of the expressivity is controlled with the bow and that is the reason of informing the model with bowing motion acquired during real performances.

In this work we are focused on Jigs, fast tunes but slower that reels that usually consist of eighth notes in a ternary time signature, with strong accents at each beat. Regarding ornaments we will concentrate on: a fingering ornament, *mordent*, thought of as a rapid single alternation between an indicated note, a note one semitone above and the indicated note again; and on a bowing ornament, *bowed triplet*, similar to a triplet but consisting of three very short and fast notes with the same pitch and different bow direction.

3. DATA ACQUISITION

The training data used in our experimental investigations are monophonic recordings of nine celtic jigs performed by a professional musician. Apart from the tempo (he was following a metronome), the musician was not given any particular instructions on how to perform the pieces.

A set of audio and bowing 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.

Audio is captured with a violin pickup and bowing with a 3D motion measuring system based on a commercial Polhemus [1] device. The main advantage compared to other expressivity models is that of having motion information. It is important in order to learn the model as well as for the alignment and segmentation of the performances with the scores.

3.1. Scores

Scores are represented as a series of notes with onset, pitch(in semitones) and duration. They represent the nominal attributes of the tune, and are used as a guide to segment the performance and find deviations from the score.

3.2. Audio acquisition

Audio is captured by means of a violin bridge pickup [2]. This way we get a signal with much lower effect of the resonances of the violin acoustic box and the room which makes segmentation much easier than for a recording from a microphone. From the captured audio stream we extract the audio perceptual features: frame-by-frame energy, fundamental frequency estimation and aperiodicity function.

3.3. 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. Two bowing descriptors, bow direction change and playing string change, are used for the segmentation.

4. DATA ANALYSIS

4.1. 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 a similar method to the one that we already described in [6] which we briefly introduce here. The procedure uses score time information in order to search for note onset/offset times around their nominal values, allowing for timing deviations up to the maximum duration of the notes involved in the transition into consideration. Phrase-starting and phrase-ending notes are treated differently, by applying a simple energy-based onset/offset detector. For score-performance alignment we use bowing data (bow speed profile), estimation of string being played, sound aperiodicity function provided by YIN [4] and sound energy envelope. Within a time window around the nominal note time values, we collect possible candidates positions of note change by detecting: (1) bow direction changes by looking for zero-crossings of the bow speed profile, (2) string changes by looking for steps in the string detection function, (3) a local maximum of the aperiodicity function, and (4) a local minima of energy envelope or energy envelope high curvature points (second derivative extrema). Then, we compute a weighted cost function (weights are set empirically) based on the position of the candidates and their nature, and the minimum of such function is considered as the note change time. See an illustration of the of the alignment procedure in fig.1. We had to manually correct some minor alignment errors due mainly to pitch changes inside a note when the performer made use of some pitch-based ornaments (see next sections).

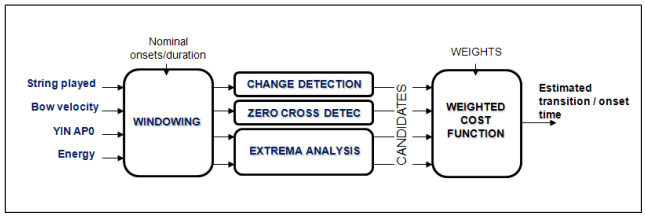

Figure 1. Schematic view of the score-performance alignment procedure

4.2. Bow direction detection

It is not enough to use detected zero-crossings of bow speed for extracting bow direction because of the possible performed *bowed triplets*. Instead, we compute bow speed histograms for each of the note segments after alignment, and get the sign of the histogram maximum as the indicator for the bow direction.

4.3. Ornaments detection

Two types of ornaments are recognized: *mordents* and *bowed triplets*. Detection of mordents is carried out after note segmentation based on based the work in [9] and can be summarized as follows: For each segmented note, pitch and aperiodicity function is calculated. Pitch curve segments with high aperiodicity or very short segments are not considered for the detection. The remaining intranote pitch curve is then quantized to semitones and notes with changes of one tone or semitone inside the note are considered to be part of an ornament. In Figure 2 a mordent is detected in a note. Bowed triplets are detected when three consecutive bow position changes occur in a short analysis frame.

[1] www.polhemus.com
[2] www.lrbaggs.com

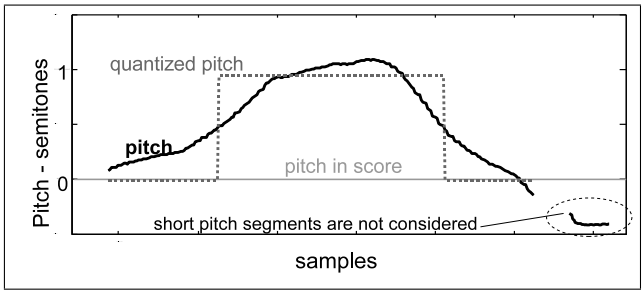

Figure 2. Mordent detection: Intranote pitch(bold) is quantized(dashed). Changes of quantized pitch respect to score pitch(thin) of one or two semitones indicate a fingering ornament inside the note. Parts without pitch have high values of aperiodicity.

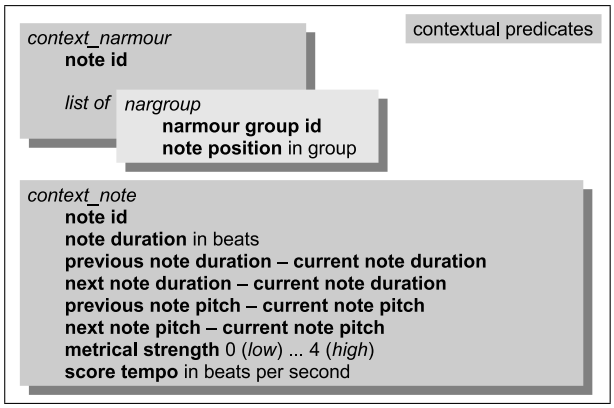

Figure 3. Contextual predicates

5. LEARNING THE EXPRESSIVE PERFORMANCE MODEL

In this section we describe our inductive approach for learning the model by applying ILP techniques. After the alignment and segmentation, scores and expressive deviations of the performance are defined in a structured way using first order logic predicates.

5.1. Data Description

The musical context of each note is defined with the following predicates (Figure 3): *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

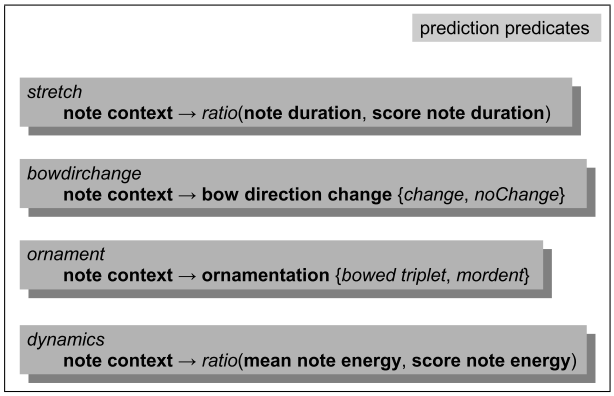

Figure 4. Induction and Prediction predicates

using 4 predicates (Figure 4): *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; *ornament* specifies whether a note is ornamented in the performance; 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.

We obtained correlation coefficients of 0.88 and 0.83 for the duration transformation and note dynamics prediction tasks, respectively and we obtained a correctly classified instances percentage of 92% and 86% for the bow direction and ornamentation prediction. These numbers were obtained by performing 10-fold cross-validation on the training data.

The induced models seem to capture accurately the expressive transformations the musician introduces in the performances. Figure 5 contrasts the note duration deviations predicted by the model and the deviations performed by the violinist. Similar results were obtained for the dynamics model.

6. EVALUATION BY SYNTHESIS

Although we give an estimation of the model performance errors, listening tests are necessary in order to asses the validity of the prediction. For this we feed a violin syn-

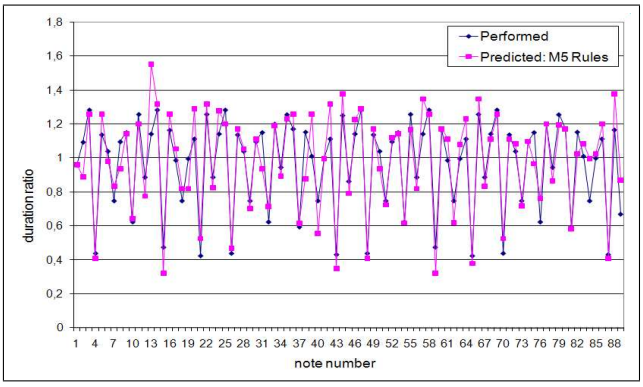

Figure 5. Note deviation ratio for a tune with 89 notes. Comparison between performed and predicted.

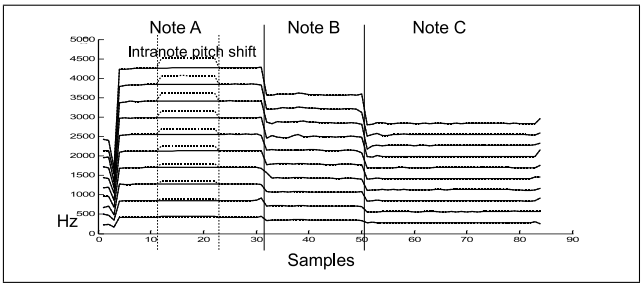

Figure 6. Mordent synthesis: Figure shows first 10 harmonic tracks of three notes. In the middle of note A a pitch shift of one semitone is done (dashed line). When resynthesizing we obtain a synthetic mordent at note A.

thesizer with the prediction results of the model. We make use of a sample-based spectral concatenative synthesizer, that uses an annotated database of samples from real performances recorded and segmented using the techniques described above. The database consist of several musical phrases combining different dynamics, durations, pitch and articulations. Note annotations include bowing information (note bow direction and whether a note is slurred or tied). Sample selection is based on a weighted Euclidean distance measure with strict matching for bow direction in a similar way as presented in [7]. After the sample selection stage, pitch shift and time stretch transformations are applied in order to match note characteristics in the model's output.

Regarding ornaments, mordents are synthesized by applying a one semitone intranote pitch shift simulating a short note (Figure6). Bowed triplets are synthesized by selecting prerecorded ones from the database depending on pitch and applying pitch shift.

7. CONCLUSIONS AND FURTHER WORK

We presented a model for expressive performances based not only on perceptual features but also informed with bowing and 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.

A small subset of perceptual and bowing features was modeled but the modeling procedure could be easily extrapolated to other ornaments, forms and styles in the future.

Although we are using high level gesture features (bowing) for this work, the potential of acquiring gesture control data with the described system is enormous. In the future we will extend the analysis of expressivity to low level features such as bow pressing force, bow velocity or bow-bridge distance.

8. 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.

9. 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] A. de Chevigne and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *JASA*, 111:4, 2002.

[5] 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.

[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] E. Maestre, A. Hazan, R. Ramirez, and A. Perez. Using concatenative synthesis for expressive performance in jazz saxophone. In *Proceedings of International Computer Music Conference*, New Orleans, 2006.

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

[9] M. Puiggros, E. Gomez, R. Ramirez, X. Serra, and R. Bresin. Automatic characterization of ornamentation from bassoon recordings for expressive synthesis. In *Proceedings of International Conference on Music Perception and Cognition*, Bologna, Italy, 2006.

[10] 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.

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