PERFORMER IDENTIFICATION IN CELTIC VIOLIN RECORDINGS

Rafael Ramirez, Alfonso Perez and Stefan Kersten
Music Technology Group
Universitat Pompeu Fabra
Ocata 1, 08003 Barcelona, Spain
Tel:+34 935422864, Fax:+34 935422202
{rafael, aperez, skersten}@iua.upf.edu

ABSTRACT

We present an approach to the task of identifying performers from their playing styles. We investigate how violinists express and communicate their view of the musical content of Celtic popular pieces and how to use this information in order to automatically identify performers. We study note-level deviations of parameters such as timing and amplitude. Our approach to performer identification consists of inducing an expressive performance model for each of the interpreters (essentially establishing a performer dependent mapping of inter-note features to a timing and amplitude expressive transformations). We present a successful performer identification case study.

1 INTRODUCTION

Music performance plays a central role in our musical culture today. Concert attendance and recording sales often reflect people's preferences for particular performers. The manipulation of sound properties such as pitch, timing, amplitude and timbre by different performers is clearly distinguishable by the listeners. Expressive music performance studies the manipulation of these sound properties in an attempt to understand expression in performances. There has been much speculation as to why performances contain expression. Hypothesis include that musical expression communicates emotions and that it clarifies musical structure, i.e. the performer shapes the music according to her own intentions

In this paper we focus on the task of identifying violin performers from their playing style using high-level descriptors extracted from single-instrument audio recordings. The identification of performers by using the expressive content in their performances raises particularly interesting questions but has nevertheless received relatively little attention in the past.

The data used in our investigations are violin audio recordings of Irish popular music performances. We use sound analysis techniques based on spectral models [15] for extracting high-level symbolic features from the recordings. In particular, for characterizing the performances used in this work, we are interested in inter-note features representing information about the music context in which expressive events occur. Once the relevant high-level information is extracted we apply machine learning techniques [9] to automatically discover regularities and expressive patterns for each performer. We use these regularities and patterns in order to identify a particular performer in a given audio recording.

The rest of the paper is organized as follows: Section 2 sets the background for the research reported here. Section 3 describes how we process the audio recordings in order to extract inter-note information. Section 4 describes our approach to performance-driven performer identification. Section 5 describes a case study on identifying performers based on their playing style and discusses the results, and finally, Section 6 presents some conclusions and indicates some areas of future research.

2 BACKGROUND

Understanding and formalizing expressive music performance is an extremely challenging problem which in the past has been studied from different perspectives, e.g. [14], [4], [2]. The main approaches to empirically studying expressive performance have been based on statistical analysis (e.g. [12]), mathematical modeling (e.g. [17]), and analysis-by-synthesis (e.g. [3]). In all these approaches, it is a person who is responsible for devising a theory or mathematical model which captures different aspects of musical expressive performance. The theory or model is later tested on real performance data in order to determine its accuracy. The majority of the research on expressive music performance has focused on the performance of musical material for which notation (i.e. a score) is available, thus providing unambiguous performance goals. Expressive performance studies have also been very much focused on (classical) piano performance in which pitch and timing measurements are simplified.

Previous research addressing expressive music performance using machine learning techniques has included a number of approaches. Lopez de Mantaras et al. [6] report on SaxEx, a performance system capable of generating expressive solo saxophone performances in Jazz. One limitation of their system is that it is incapable of explaining the predictions it makes and it is unable to handle melody alterations, e.g. ornamentations.

Ramirez et al. [11] have explored and compared diverse machine learning methods for obtaining expressive music performance models for Jazz saxophone that are capable of both generating expressive performances and explaining the expressive transformations they produce. They propose an expressive performance system based on inductive logic programming which induces a set of first order logic rules that capture expressive transformation both at an inter-note level (e.g. note duration, loudness) and at an intra-note level (e.g. note attack, sustain). Based on the theory generated by the set of rules, they implemented a melody synthesis component which generates expressive monophonic output (MIDI or audio) from inexpressive melody MIDI descriptions.

With the exception of the work by Lopez de Mantaras et al and Ramirez et al, most of the research in expressive performance using machine learning techniques has focused on classical piano music where often the tempo of the performed pieces is not constant. The works focused on classical piano have focused on global tempo and loudness transformations while we are interested in note-level tempo and loudness transformations.

Nevertheless, the use of expressive performance models, either automatically induced or manually generated, for identifying musicians has received little attention in the past. This is mainly due to two factors: (a) the high complexity of the feature extraction process that is required to characterize expressive performance, and (b) the question of how to use the information provided by an expressive performance model for the task of performance-based performer identification. To the best of our knowledge, the only group working on performance-based automatic performer identification is the group led by Gerhard Widmer. Saunders et al [13] apply string kernels to the problem of recognizing famous pianists from their playing style. The characteristics of performers playing the same piece are obtained from changes in beat-level tempo and beat-level loudness. From such characteristics, general performance alphabets can be derived, and pianists' performances can then be represented as strings. They apply both kernel partial least squares and Support Vector Machines to this data.

Stamatatos and Widmer [16] address the problem of identifying the most likely music performer, given a set of performances of the same piece by a number of skilled candidate pianists. They propose a set of very simple features for representing stylistic characteristics of a music performer that relate to a kind of 'average' performance. A database of piano performances of 22 pianists playing two pieces by Frdric Chopin is used. They propose an ensemble of simple classifiers derived by both subsampling the training set and subsampling the input features. Experiments show that the proposed features are able to quantify the differences between music performers.

3 MELODIC DESCRIPTION

First of all, we perform a spectral analysis of a portion of sound, called analysis frame, whose size is a parameter of the algorithm. This spectral analysis consists of multiplying the audio frame with an appropriate analysis window and performing a Discrete Fourier Transform (DFT) to obtain its spectrum. In this case, we use a frame width of 46 ms, an overlap factor of 50%, and a Keiser-Bessel 25dB window. Then, we compute a set of low-level descriptors for each spectrum: energy and an estimation of the fundamental frequency. From these low-level descriptors we perform a note segmentation procedure. Once the note boundaries are known, the note descriptors are computed from the low-level values. the main low-level descriptors used to characterize note-level expressive performance are instantaneous energy and fundamental frequency.

Energy computation. The energy descriptor is computed on the spectral domain, using the values of the amplitude spectrum at each analysis frame. In addition, energy is computed in different frequency bands as defined in [5], and these values are used by the algorithm for note segmentation.

Fundamental frequency estimation. For the estimation of the instantaneous fundamental frequency we use a harmonic matching model derived from the Two-Way Mismatch procedure (TWM) [7]. For each fundamental frequency candidate, mismatches between the harmonics generated and the measured partials frequencies are averaged over a fixed subset of the available partials. A weighting scheme is used to make the procedure robust to the presence of noise or absence of certain partials in the spectral data. The solution presented in [7] employs two mismatch error calculations. The first one is based on the frequency difference between each partial in the measured sequence and its nearest neighbor in the predicted sequence. The second is based on the mismatch between each harmonic in the predicted sequence and its nearest partial neighbor in the measured sequence. This two-way mismatch helps to avoid octave errors by applying a penalty for partials that are present in the measured data but are not predicted, and also for partials whose presence is predicted but which do not actually appear in the measured sequence. The TWM mismatch procedure has also the benefit that the effect of any spurious components or partial missing from the measurement can be counteracted by the presence of uncorrupted partials in the same frame.

Note segmentation is performed using a set of frame descriptors, which are energy computation in different frequency bands and fundamental frequency. Energy onsets are first detected following a band-wise algorithm that uses some psycho-acoustical knowledge [5]. In a second step, fundamental frequency transitions are also detected. Finally, both results are merged to find the note boundaries (onset and offset information).

Note descriptors. We compute note descriptors using the note boundaries and the low-level descriptors values. The low-level descriptors associated to a note segment are computed by averaging the frame values within this note segment. Pitch histograms have been used to compute the pitch note and the fundamental frequency that represents each note segment, as found in [8]. This is done to avoid taking into account mistaken frames in the fundamental frequency mean computation. First, frequency values are converted into cents, by the following formula:

$$c = 1200 \cdot \frac{\log{(\frac{f}{f_{ref}})}}{log2} \qquad (1)$$

where $f_{ref} = 8.176$ (fref is a the reference frequency of the C0). Then, we define histograms with bins of 100 cents and hop size of 5 cents and we compute the maximum of the histogram to identify the note pitch. Finally, we compute the frequency mean for all the points that belong to the histogram. The MIDI pitch is computed by quantization of this fundamental frequency mean over the frames within the note limits.

Musical Analysis. It is widely recognized that humans perform music considering a number of abstract musical structures. In order to provide an abstract structure for the recordings under study, we decided to use Narmour's theory of perception and cognition of melodies [10] to analyze the performances.

The Implication/Realization model proposed by Narmour is a theory of perception and cognition of melodies. The theory states that a melodic musical line continuously causes listeners to generate expectations of how the melody should continue. The nature of these expectations in an individual are motivated by two types of sources: innate and learned. According to Narmour, on the one hand we are all born with innate information which suggests to us how a particular melody should continue. On the other hand, learned factors are due to exposure to music throughout our lives and familiarity with musical styles and particular melodies. According to Narmour, any two consecutively perceived notes constitute a melodic interval, and if this interval is not conceived as complete, it is an implicative interval, i.e. an interval that implies a subsequent interval with certain characteristics. That is to say, some notes are more likely than others to follow the implicative interval. Two main principles recognized by Narmour concern registral direction and intervallic difference. The principle of registral direction states that small intervals imply an interval in the same registral direction (a small upward interval implies another upward interval and analogously for downward intervals), and large intervals imply a change in registral direction (a large upward interval implies a downward interval and analogously for downward intervals). The principle of intervallic difference states that a small (five semitones or less) interval implies a similarly-sized interval (plus or minus 2 semitones), and a large interval (seven semitones or more) implies a smaller interval. Based on these two principles, melodic patterns or groups can be identified that either satisfy or violate the implication as predicted by the principles. Such patterns are called structures and are labeled to denote characteristics in terms of registral direction and intervallic difference. Figure 1 shows prototypical Narmour structures. A note in a melody often belongs to more than one structure. Thus, a description of a melody as a sequence of Narmour structures consists of a list of overlapping structures. We parse each melody in the training data in order to automatically generate an implication/realization analysis of the pieces. Figure 2 shows the analysis for a fragment of a melody.

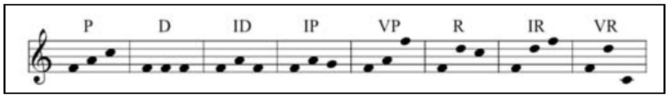

Figure 1. Prototypical Narmour Structures

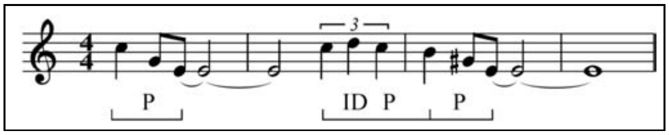

Figure 2. Narmour analysis of a melody fragment

4 PERFORMANCE-DRIVEN PERFORMER IDENTIFICATION

4.1 Note features

The note features represent both properties of the note itself and aspects of the musical context in which the note appears. Information about the note includes note pitch and note duration, while information about its melodic context includes the relative pitch and duration of the neighboring notes (i.e. previous and following notes) as well as the Narmour structures to which the note belongs. The note's Narmour structures are computed by performing the musical analysis described before. Thus, each performed note is characterized by the tuple

(Pitch, Dur, PrevPitch, PrevDur, NextPitch, NextDur, Nar1, Nar2, Nar3)

4.2 Algorithm

We are ultimately interested in obtaining a classification function F of the following form:

$$F(MelodyFragment(n1,\ldots,nk)) \longrightarrow Performers$$

where $MelodyFragment(n1,\ldots,nk)$ is the set of melody fragments composed of notes $n1,\ldots,nk$ and $Performers$ is the set of possible performers to be identified. For each performer i to be identified we induce an expressive performance model M_i predicting his/her timing and energy expressive transformations:

$$M_i(Notes) \rightarrow (PDur, PEner)$$

where $Notes$ is the set of score notes played by performer i represented by their inter-note features, i.e. each note in $Notes$ is represented by the tuple (Pitch, Dur, PrevPitch, PrevDur, NextPitch, NextDur, Nar1, Nar2, Nar3) as described before, and the vector $(PDur, PEner)$ contains the model's predictions for note duration ($PDur$) and energy ($PEner$). Once a performance model is induced for each performer P_i we apply the following algorithm:

```
F([N1,...,Nm], [P1,...,Pn])
   for each performer Pi
      Scorei = 0
   for each note Nk
      FNk = inter-note_features(Nk)
      Mi(FNk) = (PDk,PEk)
      for each performer Pi
         ScoreNKi = sqrt(((Dur(NK)-PDk)^2)
                  + ((Ener(Nk)-PEk)^2))
         Scorei = Scorei  + ScoreNKi
return Pj (j in {1,...,m}) with minimum score
```

This is, for each note in the melody fragment the classifier F computes the set of its inter-note features. Once this is done, for each note N_k and for each performer P_i, performance model M_i predicts the expected duration and energy for N_k. This prediction is based on the note's inter-note features. The score $Score_i$ for each performer i is updated by taking into account the Euclidean distance between the note's actual duration and energy and the predicted values. Finally, the performer with the lower score is returned.

Clearly, the expressive models M_i play a central role in the output of classifier F. For each performer, M_i is induced by applying Tildes 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 musical context of each note is defined by predicates $context$ and $narmour$. $context$ specifies the note features described above and $narmour$ specifies the Narmour groups to which a particular note belongs, along with its position within a particular group. Expressive deviations in the performances are encoded using predicates $stretch$ and $dynamics$. $stretch$ specifies the stretch factor of a given note with regard to its duration in the score and $dynamics$ specifies the mean energy of a given note. 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.

5 CASE STUDY

5.1 Training data

In this work we are focused on Celtic jigs, fast tunes but slower that reels, that usually consist of eighth notes in a ternary time signature, with strong accents at each beat. The training data used in our experimental investigations are monophonic recordings of nine Celtic jigs performed by two professional violinists. Apart from the tempo (they played following a metronome), the musicians were not given any particular instructions on how to perform the pieces.

5.2 Results

Initially, we evaluated the expressive performance model for each of the musicians we considered. Thus, we obtained two expressive performance models M_1 and M_2. For M_1 we obtained correlation coefficients of 0.88 and 0.83 for the duration transformation and note dynamics prediction tasks, respectively, while we obtained 0.91 and 0.85 for M_2. 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 musicians introduce in the performances. Figure 3 contrasts the note duration deviations predicted by model M_1 and the deviations performed by the violinist. Similar results were obtained for M_2.

We then proceed to evaluate the classification function F by spliting our data into a training set and a test set. We held out approximately 30% of the data as test data while the remaining 70% was used as training data (we held out 3 pieces for each violinist). When selecting the test data, we left out the same number of melody fragments per class. In order to avoid optimistic estimates of the classifier performance, we

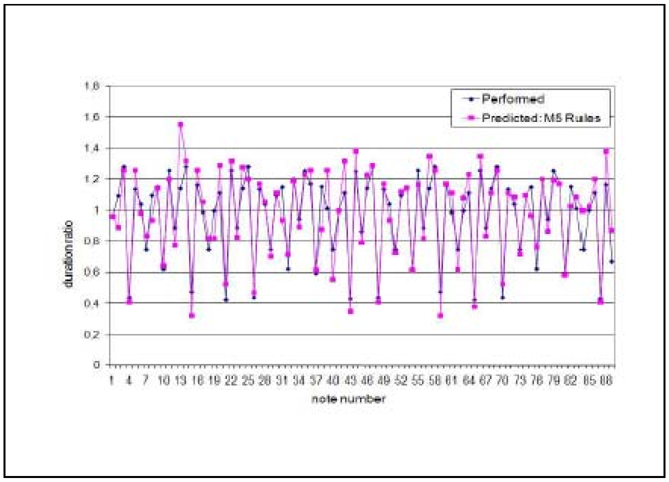

Figure 3. Note deviation ratio for a tune with 89 notes. Comparison between performed and predicted by M_1

explicitly removed from the training set all melody fragment repetitions of the hold out fragments. This is motivated by the fact that musicians are likely to perform a melody fragment and its repetition in a similar way. We tested our algorithm in each of the six test pieces (three pieces of each class) and obtained 100% accuracy (correctly classified instances percentage). This is, the six pieces in the test set were classified correctly.

6 CONCLUSION

In this paper we focused on the task of identifying performers from their playing style using note descriptors extracted from audio recordings. In particular, we concentrated in identifying violinists playing Irish popular pieces (Irish jigs). We characterized performances by representing each note in the performance by a set of inter-note features representing the context in which the note appears. We then induced an expressive performance model for each of the performers and presented a successful performer identification case study. The results obtained seem to indicate that the inter-note features presented contain sufficient information to identify the studied set of performers, and that the machine learning method explored is capable of learning performance patterns that distinguish these performers. This paper present preliminary work so there is further work in several directions. Our immediate plans are to extend our database and to test different distance measures for updating the performer scores in our algorithm. We also plan to evaluate our algorithm considering melody fragments of different size and to evaluate the predictive power of timing expressive variations relative to energy expressive variations.

7 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] Bresin, R. (2000). Virtual Visrtuosity: Studies in Automatic Music Performance. PhD Thesis, KTH, Sweden.

[3] Friberg, A.; Bresin, R.; Fryden, L.; 2000. Music from Motion: Sound Level Envelopes of Tones Expressing Human Locomotion. Journal of New Music Research 29(3): 199-210.

[4] Gabrielsson, A. (1999). The performance of Music. In D.Deutsch (Ed.), The Psychology of Music (2nd ed.) Academic Press.

[5] Klapuri, A. (1999). Sound Onset Detection by Applying Psychoacoustic Knowledge, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP.

[6] Lopez de Mantaras, R. and Arcos, J.L. (2002). AI and music, from composition to expressive performance, AI Magazine, 23-3.

[7] Maher, R.C. and Beauchamp, J.W. (1994). Fundamental frequency estimation of musical signals using a two-way mismatch procedure, Journal of the Acoustic Society of America, vol. 95 pp. 2254-2263.

[8] McNab, R.J., Smith Ll. A. and Witten I.H., (1996). Signal Processing for Melody Transcription, SIG working paper, vol. 95-22.

[9] Mitchell, T.M. (1997). Machine Learning. McGraw-Hill.

[10] Narmour, E. (1990). The Analysis and Cognition of Basic Melodic Structures: The Implication Realization Model. University of Chicago Press.

[11] Rafael Ramirez, Amaury Hazan, Esteban Maestre, Xavier Serra, A Data Mining Approach to Expressive Music Performance Modeling, in Multimedia Data mining and Knowledge Discovery, Springer.

[12] Repp, B.H. (1992). Diversity and Commonality in Music Performance: an Analysis of Timing Microstructure in Schumann's 'Traumerei'. Journal of the Acoustical Society of America 104.

[13] Saunders C., Hardoon D., Shawe-Taylor J., and Widmer G. (2004). Using String Kernels to Identify Famous Performers from their Playing Style, Proceedings of the 15th European Conference on Machine Learning (ECML'2004), Pisa, Italy.

[14] Seashore, C.E. (ed.) (1936). Objective Analysis of Music Performance. University of Iowa Press.

[15] Serra, X. and Smith, S. (1990). "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition", Computer Music Journal, Vol. 14, No. 4.

[16] Stamatatos, E. and Widmer, G. (2005). Automatic Identification of Music Performers with Learning Ensembles. Artificial Intelligence 165(1), 37-56.

[17] Todd, N. (1992). The Dynamics of Dynamics: a Model of Musical Expression. Journal of the Acoustical Society of America 91.