Technology Enhanced Learning of Expressive Music Performance

Rafael Ramirez[1]*, Fabio Ortega[1] , Sergio Giraldo[1]

[1]Music and Machine Learning Lab
Music Technology Group
Universitat Pompeu Fabra
Roc Boronat 138
08018 Barcelona, Spain

{rafael.ramirez, fabiojose.muneratti, sergio.giraldo}@upf.edu

Abstract

Learning to play music is mostly based on the master-apprentice model in which modern technologies are rarely employed and students' interaction and socialisation is often restricted to short and punctual contact with the teacher. This often makes musical learning a lonely experience, resulting in high abandonment rates. In the context of TELMI, an international project, which aims to addresses these issues by providing new multimodal interaction paradigms for music learning and to develop assistive, self-learning, real-time feedback, complementary to traditional teaching, this paper focuses on the computational modelling of expressive music performance as a tool for music learning. We record a professional violinist and apply machine learning techniques to induce an expressive model the recordings. We use this model to generate feedback on expressive aspects of arbitrary pieces to violin students.

1. Introduction

Music education requires a long learning trajectory and intensive practice. Learning to play music is mostly based on the master-apprentice model in which the teacher mainly gives verbal feedback on the performance of the student. In such a learning model, modern technologies are rarely employed and almost never go beyond audio and video recording. In addition, the student's interaction and socialisation is often restricted to short and punctual contact with the teacher followed by long periods of self-study, which often makes musical learning a lonely experience, resulting in high abandonment rates [1].

One of the most challenging skills that students must learn during their learning process is the ability to play expressively. This is usually learnt by imitating expert performers to gradually develop the own playing style. This is normally a long process where there is no general rules on how to go about it. Furthermore expressive instructions often vary considerably from teacher to teacher. TELMI is an international effort to address the challenges music instrument education poses. Concretely, TELMI aims to design and implement new multi-modal interaction paradigms and prototypes for music learning and training based on state-of-the-art audio processing and motion capture technologies, and to create a publicly available reference database of multimodal recordings with data analytics. This database, no matter how extended it is, will inevitably be non-exhaustive in the sense that it will fail to contain all possible pieces that any student may want to practice. The work described in this paper aims at extrapolating the recordings in the database by providing general expressive computational models able to generate feedback about expressive issues in arbitrary music pieces. The multi-modal recordings in the database are extended by applying machine learning techniques using database recordings as training data. We use sound analysis techniques based on spectral models [15] for extracting high-level symbolic features from the recordings. In particular, for characterising the performances used in this work, we are interested in inter-note features representing informa-

*This work has been partly sponsored by the Spanish TIN project TIMUL (TIN2013-48152-C2-2-R), the European Union Horizon 2020 research and innovation programme under grant agreement No. 688269 (TELMI project), and the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502).

tion 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.

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.

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 characterise 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 neighbour in the predicted sequence. The second is based on the mismatch between each harmonic in the predicted sequence and its nearest partial neighbour 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.

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.

4. Expressive Performance Modeling

4.1. 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 a professional violinist. Apart from the tempo (he played following a metronome), the musicians were not given any particular instructions on how to perform the pieces.

4.2. 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)

4.3. Algorithm

We apply 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.

4.4. Results

We evaluated the expressive performance model obtaining correlation coefficients of 0.88 and 0.83 for the duration transformation and note dynamics prediction tasks, respectively. These numbers were obtained by performing 10-fold cross validation on the training data. The induced model seem to capture accurately the expressive transformations the musician introduce in the performances. Figure 1 contrasts the note duration deviations predicted by the model and the deviations performed by the violinist.

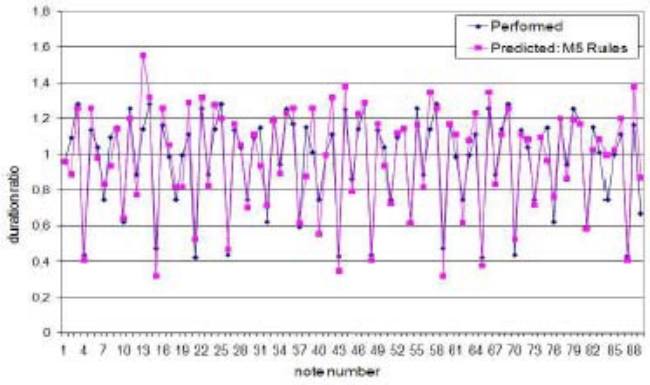

Figure 1: Note deviation ratio for a tune with 89 notes. Comparison between performed and predicted by the expressive performance model

We have implemented a prototype, which allows students to visualize the score, the expressive performance generated by the computational model, and their own performance. The prototype allows students to compare their performances with the target performance in terms of duration, and/or energy in real-time.

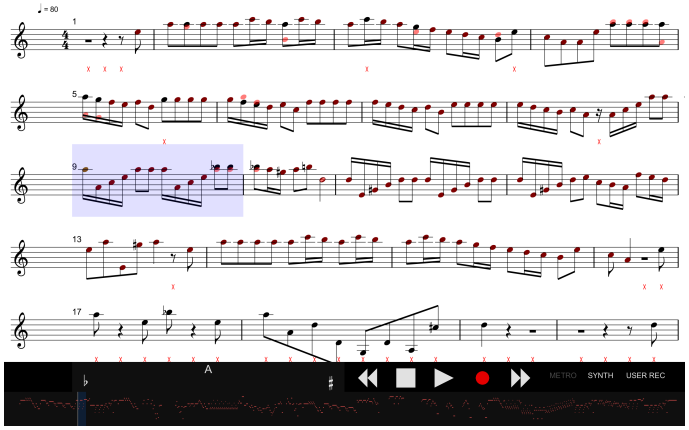

Figure 2: Visualization prototype

5. Conclusion

We applied machine learning techniques to learn computational models of music expression in violin performances. The aim is to use this models in a music learning prototype for teaching students how to play expressively. The induced models seem to capture accurately the expressive transformations the musician introduce in the performances. The implication of this work is that its outcome has the potential to contribute to the engagement of musicians in the community by making more appealing music practice and instrument training.

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. andWidmer, 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.