A Learning Scheme for Generating Expressive Music Performances of Jazz Standards

Rafael Ramirez and Amaury Hazan
Music Technology Group
Pompeu Fabra University
Ocata 1, 08003 Barcelona, Spain
{rafael,ahazan}@iua.upf.es

Abstract

We describe our approach for generating expressive music performances of monophonic Jazz melodies. It consists of three components: (a) a melodic transcription component which extracts a set of acoustic features from monophonic recordings, (b) a machine learning component which induces an expressive transformation model from the set of extracted acoustic features, and (c) a melody synthesis component which generates expressive monophonic output (MIDI or audio) from inexpressive melody descriptions using the induced expressive transformation model. In this paper we concentrate on the machine learning component, in particular, on the learning scheme we use for generating expressive audio from a score.

1 Introduction

Expressive performance is an important issue in music which has been studied from different perspectives [Gabrielsson, 1999]. The main approaches to empirically study expressive performance have been based on statistical analysis (e.g. [Repp, 1992]), mathematical modelling (e.g. [Todd, 1992]), and analysis-by-synthesis (e.g. [Friberg, 1995]). 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. Recently, there has been work on applying machine learning techniques to the study of expressive performance. Widmer [Widmer, 2002] has focused on the task of discovering general rules of expressive classical piano and recognizing famous pianists from their playing style. Lopez de Mantaras et al. [Lopez de Mantaras, 2002] reported on SaxEx, a case-based reasoning system capable of inferring a set of expressive transformations and applying them to a solo performance in Jazz. In this paper we describe an approach to investigate musical expressive performance based on inductive machine learning. In particular, we are interested in monophonic Jazz melodies performed by a saxophonist. Our work differentiates from that of Widmer in that, being focused on saxophone Jazz performances, we are interested in intra-note variations (e.g. vibrato) absent in piano, as well as melody alterations (e.g. onset deviations, ornamentations) which are normally considered performance errors in classical music. The work of Lopez de Mantaras et al. is similar to ours but they are unable to explain their predictions. The deviations and changes we consider are on note duration, note onset, note energy, and intra-note features (e.g. attack, vibrato). The study of these variations is the basis of an inductive content-based transformation tool for generating expressive performances of musical pieces. The tool can be divided into three components: a melodic transcription component, a machine learning component, and a melody synthesis component. In the following, we briefly describe each of these components.

2 Melodic description

Sound analysis and synthesis techniques based on spectral models are used for extracting high-level symbolic features from the recordings. The sound spectral model analysis techniques are based on decomposing the original signal into sinusoids plus a spectral residual. From the sinusoids of a monophonic signal it is possible to extract information on note pitch, onset, duration, attack and energy, among other high-level information. This information can be modified and the result added back to the spectral representation without loss of quality. We use the software SMSTools which is an ideal tool for preprocessing the signal and providing a high-level description of the audio recordings, as well as for generating an expressive audio according to the transformations obtained by machine learning methods.

The low-level descriptors used to characterize the melodic features of our recordings are instantaneous energy and fundamental frequency. The procedure for computing the descriptors is first to divide the audio signal into analysis frames and compute a set of low-level descriptors for each analysis frame. Then, a note segmentation is performed using low-level descriptor values. Once the note boundaries are known, the note descriptors are computed from the low-level and the fundamental frequency values (see [Gomez et al., 2003] for details about the algorithm).

3 Expressive performance knowledge induction

Data set. The training data used in our experimental investigations are monophonic recordings of three Jazz standards

(*Body and Soul*, *Once I loved* and *Like Someone in Love*) performed by a professional musician (a saxophone player) at 11 different tempos around the nominal tempo. The resulting data set is composed of 1936 performed notes.

Descriptors. In this paper, we are concerned with note-level (in particular note duration, note onset and note energy) and intra-note-level (in particular intra-note pitch and amplitude shape) expressive transformations. Each note in the training data is annotated with its corresponding deviation and a number of attributes representing both properties of the note itself and some aspects of the local context in which the note appears. Information about intrinsic properties of the note include note duration, note metrical position, and note envelope information, while information about its context include the note Narmour group(s) [Narmour, 1990], duration of previous and following notes, and extension and direction of the intervals between the note and the previous and following notes.

Machine learning techniques. In order to induce predictive models for duration ratio, onset deviation and energy, variation, we have applied machine learning techniques such as regression trees, model trees and support vector machines (for a complete comparison of the accuracy of these techniques, see [Ramirez et al., 2005]). Among these techniques, model trees is the most accurate method, and thus, we have based the machine learning component of our tool on this method. We have also induced rule-based models [Ramirez et al., 2004] to *explain* the predictions made by our tool. In order to induce a predictive model for intra-note features we have devised a learning scheme roughly described as follows:

1. apply k-means clustering to all the notes in the data set. We decided to set the number of clusters to five. This decision was taken after analyzing a large number of notes in our data set and considering that there were basically five qualitatively different types of note shapes. We characterize each note in the data set by its attack, sustain and release.
2. apply a classification algorithm (i.e. classification trees) to predict the cluster to which the note belongs. In order to train our classifier we used the descriptors described above.
3. given a note and its cluster, apply a nearest neighbor algorithm to determine the most similar note in the cluster. We use the pitch and duration of a note as the distance measure, i.e. given a note, we look in the predicted cluster for the closest note in duration and in pitch. We are particularly interested in duration and pitch because we want to minimize the loss in sound quality when transforming the selected note to a note with the required pitch and the computed duration.

Once we obtain all the notes in a score by applying the learning scheme described above, we proceed to 'glue' the obtained notes together. Finally, we apply an algorithm to obtain smooth note transitions. A sample of a melody produced by our learning scheme can be found at `www.iua.upf.es/~rramirez/promusic/demo.wav`.

4 Melody synthesis

The melody synthesis component transforms an inexpressive melody input into an expressive melody following the induced models. Given a melody score (i.e. an inexpressive description of a melody), our tool can either generate an expressive MIDI performance, or generate an expressive audio performance. In the second case, in addition to using the duration, onset and energy models for computing expressive deviations of these parameters, we apply the intra-note model to obtain the set of notes to be used to construct the audio expressive performance. In order to build the final audio performance we transform each of the obtained notes according to the computed duration, onset and energy deviations, and concatenate the transformed notes using an algorithm that optimizes the transitions between notes.

Acknowledgments

This work is supported by the Spanish TIC project ProMusic (TIC 2003-07776-C02-01). We would like to thank Emilia Gomez, Esteban Maestre and Maarten Grachten for processing the data.

References

[Friberg, 1995] Friberg, A., A Quantitative Rule System for Musical Performance. PhD Thesis, KTH, Sweden, 1995.

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

[Gomez et al., 2003] Gomez, E., Grachten, M., Amatriain, X., Arcos, J., "Melodic characterization of monophonic recordings for expressive tempo transformations", Proceedings of Stockholm Music Acoustics Conference, Stockholm, August 2003.

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

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

[Ramirez et al., 2004] Ramirez, R., Hazan, A., Gomez, E., Maestre, E., Understanding expressive transformations in saxophone jazz standards using inductive machine learning. Proceedings of Sound and Music Conference (SMC), Paris, October 2004.

[Ramirez et al., 2005] Ramirez, R., Hazan, A., Modeling Expressive Music Performance in Jazz, Clearwater Florida, May 2005.

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

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

[Widmer, 2002] Widmer, G., In Search of the Horowitz Factor: Interim Report on a Musical Discovery Project. Invited paper. In Proceedings of the 5th International Conference on Discovery Science, Lbeck, Springer-Verlag.