Modeling Expressive Music Performance in Jazz

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

Abstract

In this paper we describe a machine learning approach to one of the most challenging aspects of computer music: modeling the knowledge applied by a musician when performing a score in order to produce an expressive performance of a piece. We apply machine learning techniques to a set of monophonic recordings of Jazz standards in order to induce both rules and a numeric model for expressive performance. We implement a tool for automatic expressive performance transformations of Jazz melodies using the induced knowledge.

Introduction

Expressive performance is an important issue in music which has been studied from different perspectives (e.g. (Seashore 1936; Gabrielsson 1999; Bresin 2000)). The main approaches to empirically studying 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 1997)). 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.

In this paper we describe an approach to investigate musical expressive performance based on inductive machine learning. Instead of manually modelling expressive performance and testing the model on real musical data, we let a computer use machine learning techniques (Mitchell 1997) to automatically discover regularities and performance principles from real performance data (i.e. example performances of Jazz standards). We apply both regression and classification techniques in order to induce models of expressive performance. On the one hand, regression methods are considered to be *black-box* in the sense that it is very difficult (or impossible) to understand the predictions they produce. Black-box statistical approaches may be good at deriving predictions from data, but formulating understandable rules from the analysis of data is something entirely different from formulating predictive models from that data. On the other hand, classification methods are good at *explaining* the predictions they provide but are restricted to a set of discrete classes as prediction space. Our problem at hand is one that requires the prediction precision of regression methods for generating accurate solutions (i.e. expressive performances) but at the same time it is highly desirable to be able to explain the system predictions. Thus, we use the induced regression models to implement a tool for automatic expressive performance transformations of Jazz melodies, and the classification models to understand the principles and criteria for performing expressively a piece of music.

The rest of the paper is organized as follows: Section 2 briefly describes the melodic transcription process in the system, Section 3 describes our machine learning approach to model expressive music performance. In Section 4 related work is reported, and finally Section 5 presents some conclusions and indicates some areas of future research.

Melodic Description

Sound analysis and synthesis techniques based on spectral models (Serra & Smith 1990) 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 [1] 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 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-

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

[1] http://www.iua.upf.es/sms

level descriptor values. Once the note boundaries are known, the note descriptors are computed from the low-level and the fundamental frequency values (see (Gómez 2002) for details about the algorithm). Figure 1 is a snapshot of the SMSTools software showing the audio recording and some of the low-level descriptors extracted from it.

Modeling expressive performance knowledge in Jazz

In this section, we describe our inductive approach for learning expressive performance rules from Jazz standards performances by a skilled saxophone player. Our aim is to find note-level rules which predict, for a significant number of cases, how a particular note in a particular context should be played (e.g. longer than its nominal duration). We are aware of the fact that not all the expressive transformations regarding tempo (or any other aspect) performed by a musician can be predicted at a local note level. Musicians perform music considering a number of abstract structures (e.g. musical phrases) which makes expressive performance a multi-level phenomenon. In this context, our ultimate aim is to obtain an integrated model of expressive performance which combines note-level rules with structure-level rules. Thus, the work presented in this paper may be seen as a starting point towards this ultimate aim.

The training data used in our experimental investigations are monophonic recordings of five Jazz standards (*Body and Soul, Once I loved, Donna Lee, Like Someone in Love* and *Up Jumped Spring*) performed by a professional musician at 11 different tempos. In order to discover expressive performance regularities at different tempos we divided the recordings into three groups: nominal, slow and fast. The recordings in the nominal group are performed at the piece nominal tempo (+/- 15%) while the recordings in the slow and fast groups are respectively performed slower or faster than the ones in the nominal group. As mentioned before, sound analysis and synthesis techniques based on spectral models are used for extracting high-level symbolic features from the recordings, transforming them and synthesizing a modified recording.

After the extraction of high-level symbolic features from the recordings, each note in the training data is annotated with its corresponding class 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 the note duration and the note metrical position, while information about its context include the note Narmour group(s) (Narmour 1990) (see (Ramirez *et al*. 2004) for a desciption of the Narmour groups we use), duration of previous and following notes, and extension and direction of the intervals between the note and the previous and following notes.

In this paper, we are concerned with note-level expressive transformations (in particular transformations of note duration, onset and energy). For classification, the performance classes we are interested in are *lengthen, shorten, advance, delay, louder* and *softer*. A note is considered to belong to class *lengthen* if its performed duration is 20% or more longer that its nominal duration, e.g. its duration according to the score. Class *shorten* is defined analogously. A note is considered to be in class *advance* if its performed onset is 5% of a bar earlier (or more) than its nominal onset. Class *delay* is defined analogously. A note is considered to be in class *louder* if it is played louder than its predecessor and louder then the average level of the piece. Class *softer* is defined analogously. We have explored different possible discretization schemes. In particular we have discretized the duration, onset and energy values space in 9 classes according to the degree of transformation. For instance, we have defined 4 classes for lengthen and 4 classes for shorten (one for same) for different degrees of lengthening and shortening. In this way, we have obtained a set of finer-grain rules which, in addition to explaining expressive performances principles in more detail, may be also used to generate expressive performances. However, for generating the actual expressive performances in our system, we have induced predictive regression models (i.e. model trees) for duration ratio, onset deviation and energy variation (more on this later).

Using this data we applied the C4.5 decision tree learning algorithm (Quinlan 1993) and obtained a set of classification rules directly from the decision tree generated by the algorithm. We also applied the Apriori rule learning algorithm (Agrawal, Imieliski, & Swami 1993) to induce association rules and a sequential covering algorithm to induce first order rules. Despite the relatively small amount of training data some of the rules generated by the learning algorithms turn out to correspond to intuitive musical knowledge (e.g. Rule 1 bellow corresponds to the performer's frequent intention of accentuating a note's characteristic, in this case the shortness of the note relative to its neighboring notes). In order to illustrate the types of rules found let us consider some examples of learned note-duration rules:

RULE1: `prevdur=+3` & `nextdur=+3` & `metrstrenght=weak` ⇒ `shorten`

"If both the duration of the previous and next notes are much longer (i.e. more than three times the duration of the current note) and the current note is in a weak metrical position then shorten the current note."

RULE2: `nextdur=-2` ⇒ `lengthen`

"If the duration of the next note is considerably shorter (i.e. less than half the duration of the current note) then lengthen the current note."

RULE3: `next=silence` ⇒ `lengthen`

"if the next note is a silence then lengthen the current note."

RULE4:
```
stretch(A, B, C, shorten) :-
    succ(C, D), succ(D, E),
    context(A, E, [nargroup(p, 1)| F]),
    context(A, C, [nargroup(p, 2)| F]).
```
"Shorten a note n if it belongs to a P Narmour group in

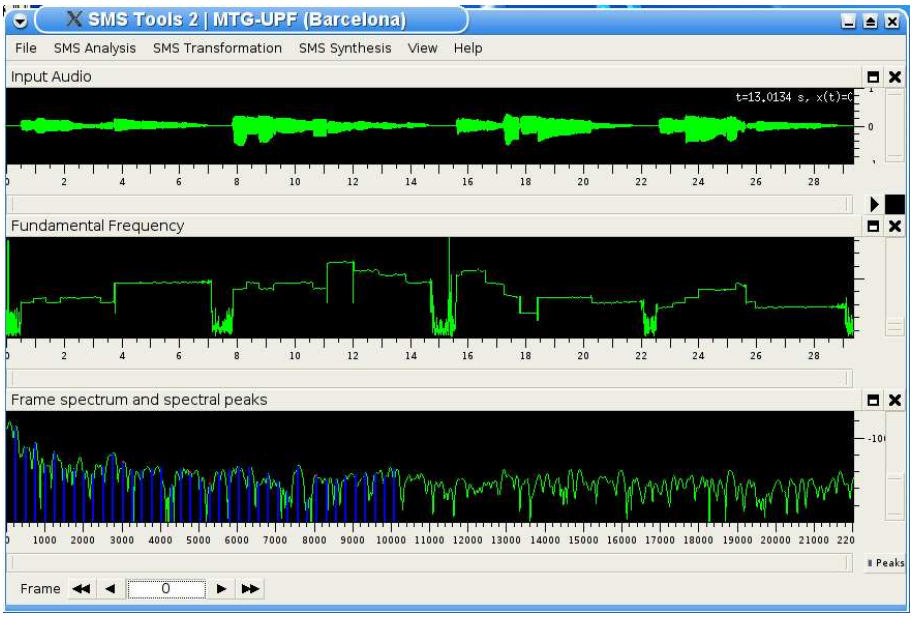

Figure 1: SMSTools showing an audio recording (top), its fundamental frequencies (middle) and spectral peaks of the first frame (bottom)

second position and if note n+2 belongs to a P Narmour group in first position"

These extremely simple rules proved to be very accurate: the first rule predicts 89%, the second rule predicts 68%, the third rule predicts 87% and the last rule predicts 96% of the relevant cases. The learning algorithm used for the rules 1 and 2 was C4.5, rule 3 was obtained by the Apriori algorithm (in this case the obtained association rules are also classification rules), and rule 4 was obtained using a sequential covering algorithm.

As mentioned before, for generating the actual expressive performances in our system, we have induced regression model trees for duration ratio, onset deviation and energy variation. We have chosen model trees as a prediction mechanism in our system because it showed the overall highest correlation coefficient among the regression methods we explored: 0.72, 0.44 and 0.67 for duration ratio, onset deviation and energy variation, respectively (we performed a 10-fold cross validation to obtain these numbers). We applied Weka's M5Rules algorithm implementation (Witten & Eibe 1999) to generate the model trees. The other methods we explored were linear regression and support vector machines with different kernels (2nd, 3rd and 4th order polynomial and radial basis).

For a detailed comparison among the regression methods we explored see Table 1 (note duration), Table 2 (note onset) and Table 3 (note energy). In Table 1, Table 2 and Table 3, C.C refers to the correlation coefficient, A.E to the relative absolute error, and S.E the root relative squared error.

Algorithm	C.C	A.E(%)	S.E(%)
Linear Regression	0.33	98.69	94.39
LMS Regression	0.29	95.22	96.60
Model Tree Regression	**0.72**	**74.89**	**69.14**
SVM Regression (1)	0.29	95.30	96.15
SVM Regression (2)	0.48	89.01	88.24
SVM Regression (3)	0.66	76.65	75.47
SVM Regression (4)	0.70	81.11	71.23

Table 1. 10-fold cross validation results for duration ratio

Algorithm	C.C	A.E(%)	S.E(%)
Linear Regression	0.17	101.12	98.41
LMS Regression	0.01	92.50	101.32
Model Tree Regression	0.43	91.51	**90.16**
SVM Regression (1)	0.14	99.92	98.88
SVM Regression (2)	0.24	**89.34**	98.18
SVM Regression (3)	0.38	95.41	92.50
SVM Regression (4)	**0.44**	94.56	90.34

Table 2. 10-fold cross validation results for onset deviation

Algorithm	C.C	A.E(%)	S.E(%)
Linear Regression	0.27	95.69	96.13
LMS Regression	0.22	87.92	108.01
Model Tree Regression	**0.67**	**66.31**	**74.31**
SVM Regression (1)	0.25	89.28	98.57
SVM Regression (2)	0.47	82.53	89.4
SVM Regression (3)	0.56	75.47	82.95
SVM Regression (4)	0.64	69.28	77.23

Table 3. 10-fold cross validation results for energy variation

Synthesis tool. We have implemented a tool which transforms an inexpressive melody input into an expressive one

following the induced model tree. The tool can either generate an expressive MIDI performance from an inexpressive MIDI description of a melody, or generate an expressive audio file from an inexpressive audio file. Figure 2 and Figure 3 show snapshots of the system applied to a MIDI performance and the system applied to an audio file, respectively. In addition to synthesising an expressive performance of a piece, the tool may be able to provide explanations for the transformations it has performed. Once a transformed note is selected requesting an explanation, the tool extracts the relevant rules generated by the different classification models and present them to the user. This explanations can be of great interest if the tool is used with padagogical purposes. Some samples of expressive performances generated by the system can be found at [2].

Related work

Previous research in learning sets of rules in a musical context has included a broad spectrum of music domains. The most related work to the research presented in this paper is the work by Widmer (Widmer 2002b; 2002a). Widmer has focused on the task of discovering general rules of expressive classical piano performance from real performance data via inductive machine learning. The performance data used for the study are MIDI recordings of 13 piano sonatas by W.A. Mozart performed by a skilled pianist. In addition to these data, the music score was also coded. The resulting substantial data consists of information about the nominal note onsets, duration, metrical information and annotations. When trained on the data the inductive rule learning algorithm named PLCG (Widmer 2001) discovered a small set of 17 quite simple classification rules (Widmer 2002b) that predict a large number of the note-level choices of the pianist. In the recordings, the tempo of the performed piece was not constant, as it was in our experiments. In fact, the tempo transformations throughout a musical piece were of special interest.

Other inductive machine learning approaches to rule learning in music and musical analysis include (Dovey 1995), (Van Baelen & De Raedt 1996), (Morales 1997) and (Igarashi, Ozaki, & Furukawa 2002). In (Dovey 1995), Dovey analyzes piano performances of Rachmaniloff pieces using inductive logic programming and extracts rules underlying them. In (Van Baelen & De Raedt 1996), Van Baelen extended Dovey's work and attempted to discover regularities that could be used to generate MIDI information derived from the musical analysis of the piece. In (Morales 1997), Morales reports research on learning counterpoint rules. The goal of the reported system is to obtain standard counterpoint rules from examples of counterpoint music pieces and basic musical knowledge from traditional music. In (Igarashi, Ozaki, & Furukawa 2002), Igarashi et al. describe the analysis of respiration during musical performance by inductive logic programming. Using a respiration sensor, respiration during cello performance was measured and rules were extracted from the data together with musical/performance knowledge such as harmonic progression and bowing direction.

Tobudic et al. (Tobudic & Widmer 2003) describe a relational instance-based approach to the problem of learning to apply expressive tempo and dynamics variations to a piece of classical music, at different levels of the phrase hierarchy. The different phrases of a piece and the relations among them are represented in first-order logic. The description of the musical scores through predicates (e.g. `contains(ph1,ph2)`) provides the background knowledge. The training examples are encoded by another predicate whose arguments encode information about the way the phrase was played by the musician. Their learning algorithm recognizes similar phrases from the training set and applies their expressive patterns to a new piece.

Lopez de Mantaras et al. (Lopez de Mantaras & Arcos 2002) report on a performance system capable of generating expressive solo performances in jazz within a case-based reasoning system. As ours, their system focuses on note onset, duration and energy. However, their system is incapable of explaining the predictions it makes.

Conclusion

This paper describes an inductive approach to learning both rules and a numeric model for expressive performance from Jazz standards recordings by a skilled saxophone player. In order to induce expressive perfromance knowledge, we have extracted a set of acoustic features from the recordings resulting in a symbolic representation of the performed pieces. We then applied both classification and regression methods to the symbolic data and information about the context in which the data appears. We used the induced regression models to implement a tool for automatic expressive performance tranformations of Jazz melodies, and we used the classification models to understand the principles and criteria for performing expressively a piece of music.

Future work: There is future work in different directions. We are currently using the timbral information of our recordings to extend our models in order to predict intra-note characteristics of an expressive performance. We plan to increase the amount of training data. Increasing the training data, extending the information in it and combining it with background musical knowledge will certainly generate a more complete model. As mentioned earlier, we intend to incorporate structure-level information to obtain an integrated model of expressive performance which combines note-level knowledge with structure-level knowledge.

References

Agrawal, R.; Imieliski, T.; and Swami, A. 1993. Mining association rules between sets of items in large databases. In *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, 207–216. ACM Press.

Bresin, R. 2000. *Virtual Visrtuosity: Studies in Automatic Music Performance.* Ph.D. Dissertation, Kungliga Tekniska hgskolan.

[2] http://www.iua.upf.es/∼rramirez/promusic

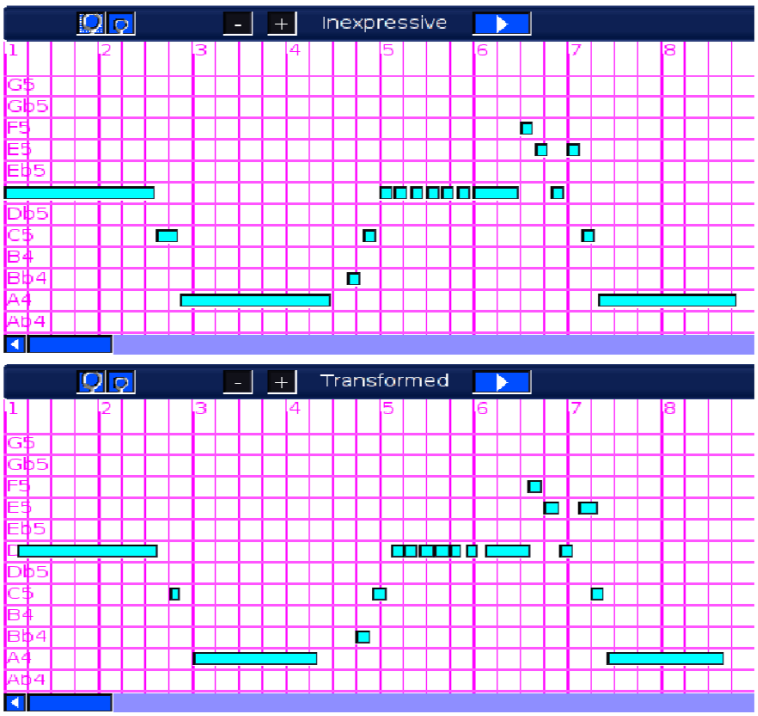

Figure 2: Expressive performance generator tool showing the inexpressive MIDI description of *Body and Soul* and the transformed expressive MIDI description

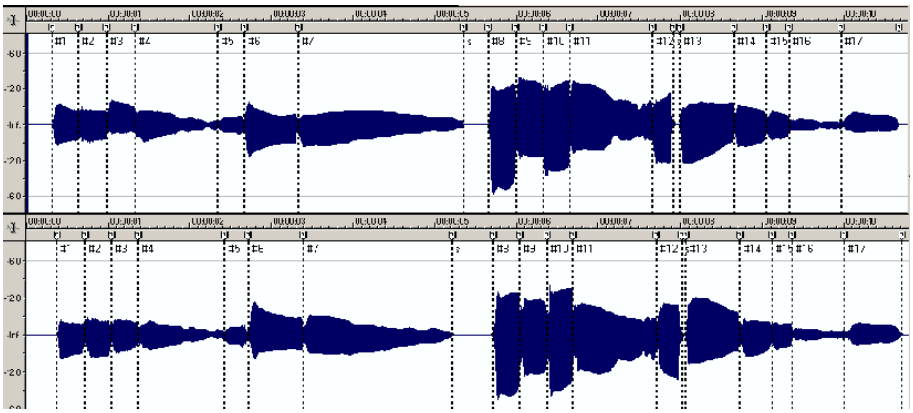

Figure 3: Expressive performance generator tool showing the inexpressive audio file of *Body and Soul* and the transformed expressive audio file

Dovey, M. 1995. Analysis of rachmaninoff's piano performances using inductive logic programming. In *Proceedings of European Conference on Machine Learning*.

Friberg, A. 1997. *A Quantitative Rule System for Musical Performance.* Ph.D. Dissertation, Kungliga Tekniska hgskolan.

Gabrielsson, A. 1999. *The performance of Music.* Academic Press.

Gómez, E. 2002. *Melodic Description of Audio Signals for Music Content Processing*. Ph.D. Dissertation, Pompeu Fabra University.

Igarashi, S.; Ozaki, T.; and Furukawa, K. 2002. Respiration reflecting musical expression: Analysis of respiration during musical performance by inductive logic programming. In *Proceedings of Second International Conference on Music and Artificial Intelligence*.

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

Mitchell, T. 1997. *Machine Learning.* McGraw-Hill.

Morales, E. 1997. Pal: A pattern-based first-order inductive system. *Machine Learning* 26.

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

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.

Ramirez, R.; Hazan, A.; Gmez, E.; and Maestre, E. 2004. Understanding expressive transformations in saxophone jazz performances using inductive machine learning. In *Proceedings of Sound and Music Computing '04*.

Repp, B. 1992. Diversity and commonality in music performance: an analysis of timing microstructure in schumann's 'traumerei'. *Journal of the Acoustic Society of America* 104.

Seashore, C. 1936. *Objective Analysis of Music Performance.* University of Iowa Press.

Serra, X., and Smith, S. 1990. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*.

Tobudic, A., and Widmer, G. 2003. Relational ibl in music with a new structural similarity measure. In *Proceedings of the International Conference on Inductive Logic Programming*.

Todd, N. 1992. The dynamics of dynamics: a model of musical expression. *Journal of the Acoustic Society of America* 91.

Van Baelen, E., and De Raedt, L. 1996. Analysis and prediction of piano performances using inductive logic programming. In *Proceedings of International Conference in Inductive Logic Programming*.

Widmer, G. 2001. Discovering strong principles of expressive music performance with the plcg rule learning strategy. In *Proceedings of the 12th European Conference on Machine Learning (ECML'01)*.

Widmer, G. 2002a. In search of the horowitz factor: Interim report on a musical discovery project. In *Proceedings of the 5th International Conference on Discovery Science (DS'02)*.

Widmer, G. 2002b. Machine discoveries: A few simple, robust local expression principles. *Computer Music Journal*.

Witten, I., and Eibe, F. 1999. *Data Mining, Practical Machine Learning Tools and Techniques with Java Implementation*. Morgan Kaufmann Publishers Inc.