Automatic Performer Identification in Commercial Monophonic Jazz Performances

Rafael Ramirez, Esteban Maestre, Xavier Serra
Information Systems and Telecommunications Department
Pompeu Fabra University
Tanger 122, 080018 Barcelona, Spain
Tel:+34 935421365, Fax:+34 935422202
rafael.ramirez@upf.edu, emaestre@upf.edu, xserra@upf.edu

Abstract

We present a pattern recognition approach to the task of identifying performers from their interpretative styles. We investigate how professional musicians express their view of the musical content of musical pieces and how to use this information in order to automatically identify performers. We apply sound analysis techniques based on spectral models for extracting deviation patterns of parameters such as pitch, timing, amplitude and timbre characterising both the internal structure of notes and the musical context in which they appear. We describe successful performer identification case studies involving monophonic audio recordings of both score-guided and commercial improvised performances.

1. Introduction

Performers manipulate sound properties such as pitch, timing, amplitude and timbre. These manipulations are clearly distinguishable by the listeners and often are reflected in concert attendance and recording sales. Expressive music performance research (for an overview see

(Gabrielsson, 1999, 2003)) investigates the manipulation of these sound properties in an attempt to understand and recreate expression in performances.

While in most forms of western classical music performers strictly follow a score specification of a piece, in other music genres such as Jazz, musicians are often encouraged to deviate significantly from the score (if the score exists at all) by adding notes, embellishments and rhythm variation. This is often referred as *improvisation*. In the past, expressive performance research has investigated manipulations of different sound properties in score-driven performances (e.g. (Widmer, 2002)) as well as different types of deviations from the score in popular music (e.g. (Lopez de Mantaras, 2002) , (Ramirez, 2008)).

In this paper we concentrate on automatic performer identification based on expressive content extracted from monophonic audio recordings (i.e. recordings of one instrument playing one note at a time). In particular, we investigate performer identification in both score-driven and improvisation saxophone jazz recordings. Expressive-content based performer identification raises particularly interesting questions but has nevertheless received relatively little attention in the past. Given the current capabilities of current audio analysis systems, we believe that identification of performers based on their playing styles is a promising research topic in music information retrieval. This work is based on our previous work on expressive performance modelling (Ramirez, 2006) (Ramirez, 2008).

The rest of the paper is organized as follows: Section 2 sets the background for our research reported here. Section 3 describes how we process the audio recordings in order to extract information about both the internal structure of notes (i.e. intra-note information) and the musical context in which they appear (i.e. inter-note information). Section 4 describes our approach to performance-driven performer identification. Section 5 describes two case studies on identifying performers based on their playing styles and discusses the results, and finally, Section 6 presents some conclusions and indicates some areas of future research.

2. Background

Expressive music performance studies the manipulation of sound properties such as pitch, timing, amplitude and timbre in an attempt to understand and recreate expression in performances. There has been much speculation as to *why* performances contain expression. Hypothesis include that musical expression communicates emotions (Juslin, 2001) and that it clarifies musical structure (Kendall, 1990), i.e. the performer shapes the music according to her own intensions (Apel, 1972). In any case, understanding and formalizing expressive music performance is an extremely challenging problem which in the past has been studied from different perspectives (for an overview see (Gabrielsson, 1999, 2003)). The main approaches to empirically studying expressive performance have been based on statistical analysis (e.g. (Repp, 1992)), mathematical modeling (e.g. (Todd, 1992)), and analysis-by-synthesis (e.g. (Friberg, 1998)). 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. More recently, machine learning techniques (Mitchell, 1997) have been applied in order to automatically induce an expressive performance model from a training set (e.g. Widmer, 2004). The majority of the research on expressive music performance, either empirical or machine-learning-based, has focused on classical piano for which notation (i.e. a score) is available.

2.1 Empirical expressive performance research

There are a number of approaches which address expressive performance without using machine learning techniques. One of the first attempts to provide a computer system with musical expressiveness is that of Johnson (Johnson, 1992). Johnson developed a rule-based expert system to determine expressive tempo and articulation for Bach's fugues from *the Well-Tempered Clavier*. The rules were obtained from two expert performers.

A long-term effort in expressive performance modeling is the work of the KTH group (Bresin, 2002), (Friberg, 2000, 2006). Their *Director Musices* system incorporates rules for tempo, dynamic and articulation transformations. The rules are obtained from both theoretical musical knowledge, and experimentally by using an analysis-by-synthesis manual approach. The rules are divided into *differentiation rules* which enhance the differences between scale tones, *grouping rules* which specify what tones belong together, and *ensemble rules* which synchronize the voices in an ensemble.

Canazza et al. (Canazza, 1997) developed a system to analyze the relationship between the musician's expressive intentions and her performance. The analysis reveals two expressive dimensions, one related to loudness (dynamics), and another one related to timing (rubato).

Dannenberg et al (Dannenberg, 1998) investigated the trumpet articulation transformations using (manually generated) rules. They developed a trumpet synthesizer which combines a physical model with an expressive performance model. The performance model generates control information for the physical model using a set of rules manually extracted from the analysis of a collection of performance recordings.

2.2 Machine-learning-based expressive performance research

Previous research addressing expressive music performance using machine learning techniques has included a number of approaches. Lopez de Mantaras et al. (Lopez de Mantaras, 2002) report on SaxEx, a performance system capable of generating expressive solo saxophone performances in Jazz. Their system is based on case-based reasoning, a type of analogical reasoning where problems are solved by reusing the solutions of similar, previously solved problems. In order to generate expressive solo performances, the case-based reasoning system retrieves from a memory containing expressive interpretations, those notes that are *similar* to the input inexpressive notes. The case memory contains information about metrical strength, note duration, and so on, and uses this information to retrieve the appropriate notes. 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. (Ramirez, 2006) 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, e.g. (Dovey, 1995), (Baelen, 1996), (Widmer, 2001), (Tobudic, 2003), where often the tempo of the performed pieces is not constant. Thus, these works focus on *global* tempo and loudness transformations.

2.3 Expressive performance research and performer identification

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. (Saunders, 2004) 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 (Stamatatos, 2005) 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 Frédéric 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.

2.4 Current Study

This paper describes a machine learning approach to investigate how skilled musicians express their view of the emotional content of musical pieces and how to use this information in order to automatically distinguish among performers. We extract features from monophonic audio recordings at two levels: intra-note and inter-note. The intra-note features represent the internal structure of a note (e.g. note attack), while the inter-note features represent aspects of the musical context in which the note appears (e.g. pitch interval with previous note). In particular, we study deviations of parameters such as pitch, timing, amplitude and timbre both at an inter-note-level and at an intra-note-level. This is, we analyze the pitch, timing (onset and duration), amplitude (energy mean) and timbre of individual notes, as well as the timing and amplitude of individual intra-note events. We focus on saxophone audio performances where timing and pitch measurements present a challenge compared to e.g. MIDI piano performances.

Roughly, the basic idea of our approach to performer identification is to establish a performer-dependent mapping from inter-note features to a repertoire of inflections characterized by intra-note features. As an analogy, the inter-note features may be seen as a literary text, while the repertoire of inflections (i.e. the intra-note features) is like a typeface or style of handwriting that different performers use to render the text in different ways. Our approach to performer identification is motivated by our pervious work (Ramirez, 2005b) on expressive music performance synthesis. In (Ramirez, 2005b) we consider a set of inflections (characterized by intra-note features) and use the note musical context (characterized by inter-note features) in order to predict the type of inflection to be used in that context. We use particular instances, i.e. audio samples, of the type of inflection predicted to synthesize expressive performances from inexpressive score descriptions. It is clear that by using a particular performer's samples the synthesized pieces 'sound' like played by that performer. Thus, it seems reasonable to apply the inverse process for performer identification.

3 Audio Analysis

In this section, we describe how we extract a description of a performed melody for monophonic recordings (for a comparison of the method reported here and other methods see (Gomez, 2003)). For each note in the recordings, we are interested in obtaining a set of descriptors characterising the internal structure of the note (i.e. attack level, sustain duration, sustain slope, amount of legato with the previous note, amount of legato with the following note, mean energy, spectral centroid and spectral tilt) and a set of descriptors charaterising the musical context in which the note appears (i.e. relative pitch and duration of the neighboring notes as well as the musical structures to which the note belongs). Audio analysis data obtained with these methods has already been used for expressive performance rule induction (Ramirez, 2006), intra-note feature prediction (Ramirez, 2005), and genetic algorithms-based expressive performance modelling (Ramirez, 2008).

3.1 Description scheme

We extract descriptors related to different temporal scales:

- Some features are defined as instantaneous or related to an analysis frame, such as energy, fundamental frequency, spectral centroid and spectral tilt.

- We also obtain intra-note/inter-note segment features, i.e. descriptors attached to a certain intra-note segment (*attack*, *sustain* and *release* segments, considering the classical ADSR model (Bernstein, 1976)) or transition segment. After observing the shape of the energy envelope of recorded notes, we realized that most of the notes did not present a clear decay segment but a fairly constant slope sustain segment. Thus, we decided not to consider decay and sustain segments separately, but just a linear sustain segment with variable slope.
- Note features or descriptors attached to a certain note are also extracted.

We considered that the proposed features constitute a simple but concise scheme for representing the typical expressive nuances in which we are interested.

3.2 Extraction of Inter-note Features

The first step once the low-level descriptors have been extracted for each frame is to get a melodic description of the audio phrases consisting on the exact onset and duration of notes, and the corresponding MIDI equivalent pitch. We base our melody transcription on the extraction of two different onset streams, the first based on energy, and the second based on fundamental frequency. Energy onsets are first detected following a band-wise algorithm that uses some psycho-acoustical knowledge (Klapuri 1999). In a second step, fundamental frequency transitions are also detected. Finally, both results are merged to find note boundaries (see Figure 1). 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 are used to compute the pitch note of each note segment, as found in (McNab, 1996). This is done to avoid taking into account mistaken frames in the fundamental frequency mean computation. First, frequency values are converted into cents. 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. An extended explanation of the methods we use for melodic description can be found in (Gomez et al. 2003).

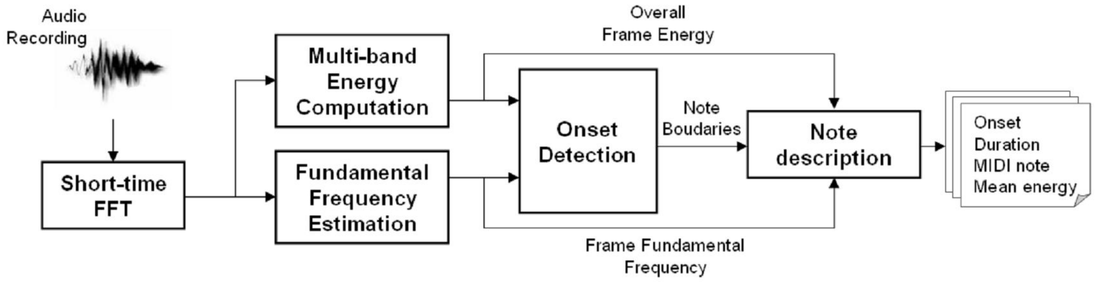

Figure 1. Schematic view of the melodic description process. Note onsets are extracted based on the study of energy and fundamental frequency**.**

Note Transitions. For characterizing note detachment, we also extract some features of the note-to-note transitions describing how two notes are detached. For two consecutive notes, we consider the transition segment starting at the first note's release and finishing at the attack of the following one. Both the energy envelope and the fundamental frequency contour (schematically represented by E_{XX} *and* f_0 in Figure 1) during transitions are studied in order to extract descriptors related to articulation. We measure the energy envelope minimum position t_c (see also Figure 2) with respect to the transition duration as (1). This descriptor has proven useful when reconstructing amplitude envelopes during transitions.

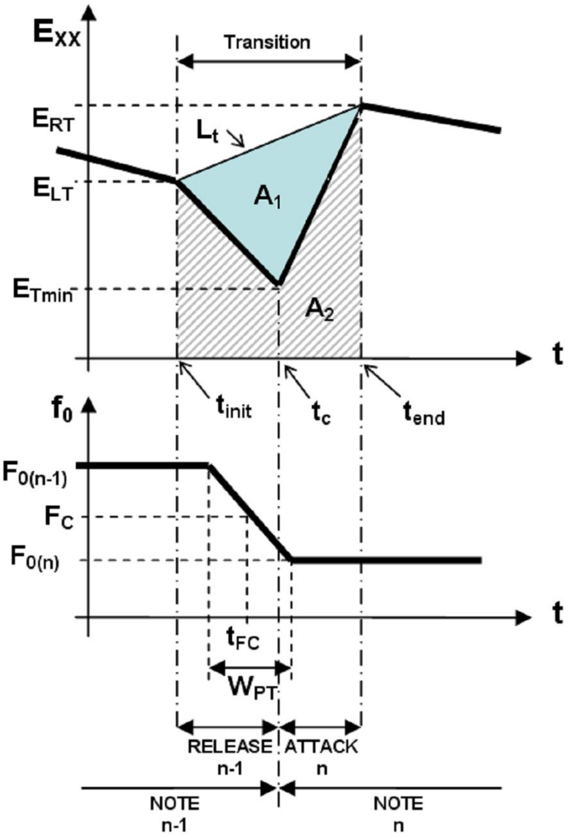

Figure 2: Schematic view of the transition segment characterization

$$E_{TPOS_{min}} = \frac{t_c}{t_{end} - t_{init}} \tag{1}$$

We compute a legato descriptor as described next. First, we join start and end points on the energy envelope contour by means of a line L_t representing the smoothest case of detachment. Then, we compute both the area A_2 below energy envelope and the area A_1 between energy envelope and the joining line L_t and define our legato descriptor as shown in (2). The relevance of this descriptor was assessed in (Maestre, 2005).

$$LEG = \frac{A_1}{A_1 + A_2} = \frac{\int_{t_{init}}^{t_{end}} \left(L_t(t) - E_{XX}(t)\right) dt}{\int_{t_{init}}^{t_{end}} L_t(t)\, dt} \tag{2}$$

Musical Analysis. It is widely recognized that expressive performance is a multi-level phenomenon and that humans perform music considering a number of abstract musical structures. After having computed the note descriptors as above, and as a first step towards providing an abstract structure for the recordings under study, we decided to use Narmour's theory of perception and cognition of melodies (Narmour 1990), (Narmour, 1991) to analyse 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. 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 3 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 4 shows the analysis for a fragment of a melody.

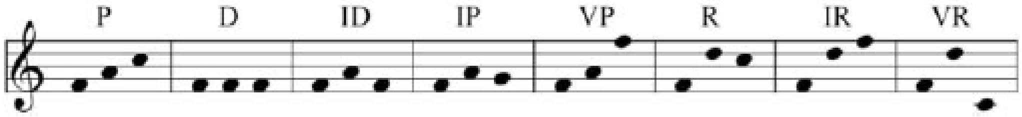

Fig. 3 Prototypical Narmour Structures

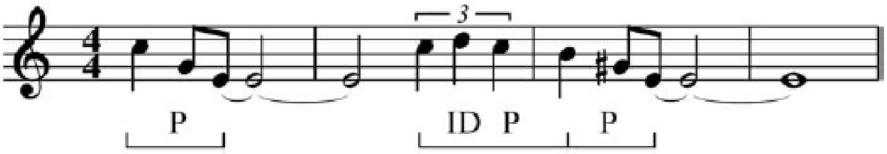

Fig. 4 Narmour analysis of a melody fragment

3.3 Extraction of Intra-note Features

Once we segment the audio signal into notes, we perform a characterization of each of the notes in terms of its internal features. We described in detail and evaluated the procedure for carrying out intra-note segmentation in (Maestre, 2005).

Intra-note segmentation. The intra-note segmentation method is based on the study of the energy envelope contour of the note. Once onsets and offsets are located, we study the instantaneous energy values of the analysis frames corresponding to each note. This study is carried out by analyzing the envelope curvature and characterizing its shape, in order to estimate the limits of the intra-note segments.

When observing the note energy envelopes from the saxophone recordings, we identify that there are usually three segments (attack, sustain and release (Bernstein, 1976)) needed to conform a description that fits the model schematically represented in Figure 5. We discarded the decay segment due to the general characteristics of the notes within the performances.

In order to extract these three characteristic segments, we study the smoothed derivatives in a similar way as presented in (Jenssen, 1999), where partial amplitude envelopes are modeled for isolated sounds. The main difference is that we analyze the notes in their musical context, rather than isolated. In addition, only three linear segments are considered. Moreover, instead of studying the contribution of all the partials, we obtain general intensity information from the total energy envelope characteristic. The procedure is carried out as follows.

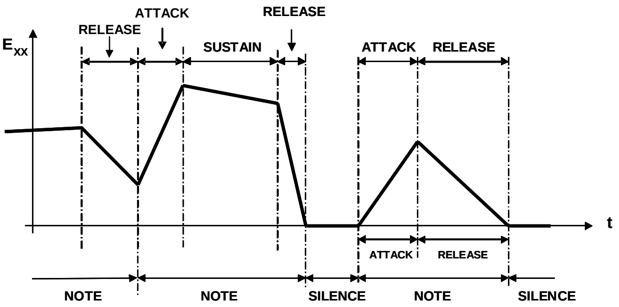

Fig. 5. Schematic view of the proposed energy envelope-based intra-note segmentation.

Considering the energy envelope as a differentiable function over time, the points of maximum curvature can be considered as the local maximum variations of the first derivative of the signal energy (second derivative extremes), that is, the local maxima or minima of the second derivative.

Due to the characteristics of the audio signal, the energy envelope must be previously smoothed by low-pass filtering, since there are typically too many second derivative extremes. The low-pass filtering is carried out by means of a variable-width Gaussian convolution. Several smoothing steps are carried out in order to find a good cut-off frequency of the smoothing filter. The smoothed envelope should not differ much to the original one to avoid loss of localization due to the filtering effect. Thus, for each smoothing step, the error e_m at smoothing step m between original and current envelope is computed. This is carried out by means of (3), where N is the length of the envelope in frames, env is the original envelope and env_m is the smoothed envelope at step m.

$$e_m = \frac{1}{N}\sum_{k=1}^{N}\frac{\left|env(k) - env_m(k)\right|}{\overline{env}} \quad (3)$$

Starting from a low cut-off frequency f_{0init}, this frequency is increased each smoothing step until the error e_m gets lower than a certain threshold e_{th}., empirically selected. Then, we compute the three first derivatives of the last smoothed envelope. Frame positions and corresponding y-values of second derivative extremes are stored. Afterwards, these characteristic points are sorted by the second derivative modulus, and the *n* highest positions are selected to build up the set of characteristic points *F*. Of course, when the total number of third derivative zero-crossings is less than *n*, the set is *F* shortened.

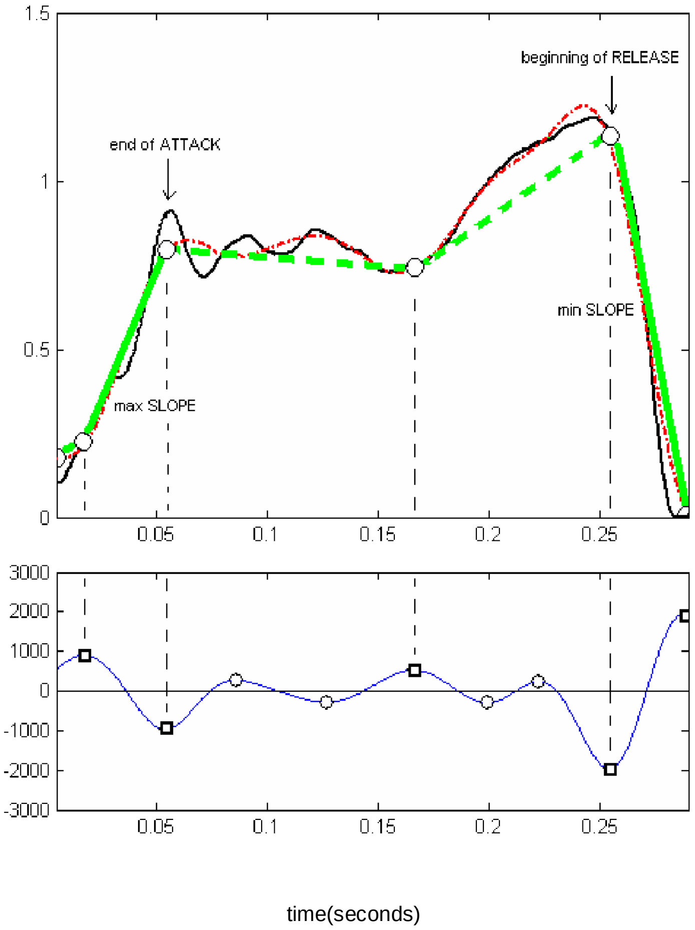

Fig. 6. Top figure: original (solid line) and smoothed (red dashed line) envelopes of a sax note for a value of e_{th}=0.05. Bottom figure: selected characteristic points are denoted with a square within extremes of the second derivative of the smoothed envelope (red dashed line).

Both note onset and offset are added as characteristic points to the set *F*. The slope defined by each pair of consecutive characteristic points on the envelope is computed (4), where *i* and *j* denote frame positions. A minimum slope duration (measured in frames) ∆*fr* is defined relative to the note duration as the five per cent of the note length N for excluding the possible too high valued slopes near the note limits.

$$\forall i, j \in F \text{ such as } i \leq j + \Delta fr, s_{i,j} = \frac{env_m(j) - env_m(i)}{j - i} \quad (4)$$

Finally, the two pairs of points defining, respectively, the most positive and most negative slope values from the remaining slopes after discarding are extracted. The end of the attack segment f_{AE} is defined as the frame position corresponding to second point of the maximum slope, while the start of the release segment position f_{RB} is defined as the first point of the minimum slope. This is stated in (5) and (6) and depicted in figure 6.

$$s_M = s_{i_M, j_M} = \max(s_{i,j}) \ , \quad f_{AE} = j_M \quad (5)$$

$$s_m = s_{i_m, j_m} = \min(s_{i,j}) \ , \quad f_{RB} = i_m \quad (6)$$

The attack is defined as the segment between the note onset and the end of the most positive of the computed slopes, while the release segment is defined as the segment between the start of the most negative of the computed slopes and the note offset. Sustain is restricted to the remaining segment. When the end of attack and the start of release limits of a note coincide, it is considered that the note does not have a sustain segment.

Intra-note segment characterization. Once we have found the intra-note segment limits, we describe each one by its duration (absolute and relative to note duration), start and end times, initial and final energy values (absolute and relative to note maximum) and slope. For the stable part of each note (sustain segment), we extract an averaged spectral centroid and spectral tilt in order to have timbral descriptors related to the brightness of a particular execution. We compute the spectral centroid as the frequency bin corresponding to the barycenter of the spectrum, expressed as (7), where *fft* is the fast fourier transform of a frame, *N* is the size of the fast Fourier transform, and *k* is the bin index. For the spectral tilt, we perform a linear regression of the logarithmic spectral envelope between 2kHz and 6kHz, and get the slope expressed in dB/Hz.

$$SC = \frac{\sum_{k=1}^{N} k \left| fft(k) \right|}{\sum_{k=1}^{N} \left| fft(k) \right|} \qquad (7)$$

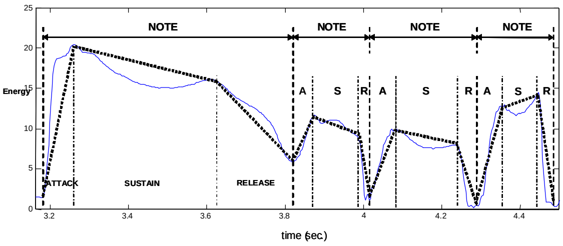

Fig. 7. Energy envelope and its linear approximation of a real excerpt with intra-note segment limits marked.

4. Performance-driven Performer Identification

In this section, we describe our approach to the problem of identifying performers from their playing style. Our approach consists of modelling each performer's playing style so that when presented with a new performance, an informed judgement may be made as regards which of the performer's models most closely matches the new performance. We introduce the different note descriptors we use to characterize audio performances (computed as described in the previous section), as well as the different pattern recognition techniques involved in the system.

4.1 Note Descriptors

We characterize each performed note by the following two sets of features:

- *Intra-note features*. The intra-note features represent the internal structure of a note which is specified as intra-note characteristics of the audio signal. The set of intra-note features we have included in the research reported here are the note's attack level, sustain duration, sustain slope, amount of legato with the previous note, amount of legato with the following note, mean energy, spectral centroid and spectral tilt. This is, each performed note is characterized by the tuple

  (*AtackLev, SustDur, SustSlo, LegLeft, LegRight, EnergyM, SpecCen, SpecTilt*)

- *Inter-note features*. The inter-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 (*Nar1, Nar2* and *Nar3* denote the three Narmour structures with the considered note in position 1, 2 and 3, respectively). The note's Narmour structures are computed by performing the musical analysis described in Section 3.2. Thus, each performed note is contextually characterized by the tuple

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

4.3 Algorithm

A central question to be asked before attempting to build a system to automatically identify a musician by his or her playing style is: how is this task performed by a music expert? The answer depends on the music genre and the instrument being played. Clearly, timing, dynamics and timbre aspects of the performance are all important for identifying particular performers, the relative importance of each aspect is not straightforward. On the one side of the instruments spectrum we could identify the singing voice in which the timbre aspect of the performance (i.e.

the singer's particular voice timbre) is of paramount importance for identifying a particular singer. On the other side of the spectrum we could identify the piano for which timing and dynamics are the most important cues to identify a particular performer while the timbre is of almost no importance. In the case of saxophone we conjecture that most of the cues for performer identification come from the timbre of the notes performed by the saxophonist. That is to say, while timing information is certainly important and is useful to identify a particular musician most of the information relevant for identifying a performer is the timbre characteristics of the performed notes. In this respect, the saxophone is similar to the singing voice in which most of the information relevant for identifying a singer is simply his or her voice's timbre. Thus, the algorithm to identify performers from their playing style reported in this paper aims to detect patterns of notes based on their timbre content. Roughly, the algorithm consists of generating a performance alphabet by clustering similar (in terms of timbre) individual notes, inducing for each performer a classifier which maps a note and its musical context to a symbol in the performance alphabet (i.e. a cluster), and given an audio fragment identify the performer as the one whose classifier predicts best the performed fragment. More formally, we are ultimately interested in obtaining a classifier *MC* of the following form:

$$MC(MelodyFragment(n_1,\ldots,n_k)) \rightarrow Performers$$

where *MelodyFragment*($n_1,\ldots,n_k$) is the set of melody fragments composed of notes $n_1,\ldots,n_k$ and *Performers* is the set of possible saxophonists to be identified. For each performer *i* to be identified we trained another classifier CL_i of the following form:

$$CL_i(CNote) \rightarrow AlphabetSymbol$$

where *CNote* is the set of notes played by performer *i* represented by their inter-note features, i.e. each note in *Note* is represented by the tuple (*Pitch, Dur, PrevPitch, PrevDur, NextPitch, NextDur, Nar1, Nar2, Nar3*) as described before, and *AlphabetSymbol* is the set of clusters generated by clustering all the notes performed (by all performers) using their intra-note features.

In order to obtain the classifiers *MC* and *CL_i* we use and explore several machine learning techniques. The machine learning techniques considered in this paper are the following:

- *K-means Clustering.* Clustering techniques apply when there is no class to be predicted but rather when the instances are to be divided into natural groups. In k-means clustering (k is the number of clusters), k points are chosen at random as cluster centers, each instance is assigned to the nearest cluster center, for each cluster a new cluster center is computed by averaging over all instances in the cluster, and the whole process is repeated with the new cluster centers. Iteration continues until the same instances are assigned to each cluster in consecutive rounds. In this paper, we apply *fuzzy k-means clustering* where Instances can belong to several clusters with different 'degrees of membership'.

- *Decision Trees*. A decision tree classifier recursively constructs a tree by selecting at each node the most relevant attribute. This process gradually splits up the training set into subsets until all instances at a node have the same classification. The selection of the most relevant attribute at each node is based on the *information gain* associated with each node of the tree (and corresponding set of instances). We have applied the decision tree building algorithm C4.5 (Quinlan, 1993).

- *Support Vector Machines (SVM)*. SVM (Cristiani, 2000) take great advantage of using a non linear attribute mapping that allows them to be able to predict non-linear models (though they remain linear in a higher dimension space). Thus, they provide a flexible prediction, but with a higher computational cost necessary to perform all the computations in the higher dimensional space. The classification accuracy of SVM largely depends on the choice of the kernel evaluation function and the parameters which control the amount to which deviations are tolerated (denoted by epsilon). In this paper we have explored SVM with linear and polynomial kernels (2nd, 3rd and 4th order).

- *Artificial Neural Networks (ANN).*ANN learning methods provide a robust approach to approximating a target function. In this paper we apply a gradient descent back propagation algorithm (Chauvin, 1995) to tune the neural network parameters to best fit the fMRI training set. The back propagation algorithm learns the weights for a multi layer network, given a network with a fixed set of units and interconnections. We set the momentum applied to the weights during updating to 0.2 and the learning rate (the amount the weights are updated) to 0.3. We use a fully-connected multi layer neural network with one hidden layer (one input neuron for each attribute and one output neuron for each class).

- *Lazy Methods.* Lazy Methods are based on the notion of lazy learning which subsumes a family of algorithms that store the complete set of given (classified) examples of an underlying example language and delay all further calculations until requests for classifying yet unseen instances are received. In this paper we have explored the *k*-Nearest Neighbor (*k*-NN) algorithm (with $k \in \{1,2,3,4,7\}$) which is capable of handling noisy data well if the training set has an acceptable size. However, *k*-NN does not behave well in the presence of irrelevant attributes.

- *Ensemble Methods.* One obvious approach to making more reliable decisions is to combine the output of several different models. In this paper we explore the use of methods for combining models (called *ensemble* methods) generated by machine learning. In particular, we have explored *voting, stacking, bagging* and *boosting*. In many cases they have proved to increase predictive performance over a single model. In the voting method, a set of *n* different classifiers are trained on the same training data using different learning algorithms (in this paper we applied decision trees, SVM, ANN, and 1-NN), and prediction is performed by allowing all *n* classifiers to 'vote' on class prediction; the final prediction is the class that gets the most votes. Stacking train *n* learning algorithms (here we applied decision trees, SVM, ANN, and 1-NN) in the same training data and train another learning algorithm, the 'meta-learner', (we applied decision trees) to learn to predict the class from the predictions of the base learners. Bagging draws n bootstrap samples from the training data, trains a given learning algorithm (here we consider decision trees) on each of these *n* samples (producing *n* classifiers) and predicts by simple voting of all *n* classifiers. Boosting generates a series of classifiers using the same learning algorithm (here we applied decision trees) but differently weighted examples from the same training set, and predicts by weighted majority vote (weighted by accuracy) of all *n* classifiers.

We segmented all the recorded pieces into audio segments representing musical phrases. Given an audio fragment denoted by a list of notes [$N_1,\ldots,N_m$] and a set of possible performers denoted by a list of performers [$P_1,\ldots,P_n$], classifier *MC* identifies the performer as follows:

```
$MC([N_1,\ldots,N_m], [P_1,\ldots,P_n])$
    for each performer $P_i$
        $\mathrm{Score}_i = 0$
    for each note $N_k$
        $PN_k$ = intra-note_features($N_k$)
        $CN_k$ = inter-note_features($N_k$)
        $(X_{k1},\ldots,X_{kq})$ = cluster_membership($PN_k$)
        for each performer $P_i$
            $\mathrm{Cluster}^{i,k}=CL_i(CN_k)$
            $Score_i = Score_i + X_{\mathrm{Cluster}^{i,k}}$
return $P_M$ such that $\mathrm{Score}_M = \max(Score_1,\ldots,Score_n)$
```

This is, for each note in the melody fragment the classifier *MC* computes the set of its intra-note features, the set of its inter-note features and, based on the note's intra-note features, the cluster membership of the note for each of the clusters ($X_1,\ldots,X_q$ are the cluster membership for clusters *1,…,q*, respectively). Once this is done, for each performer P_i its trained classifier $CL_i(PN)$ predicts a cluster representing the expected type of note the performer would have played in that musical context. 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 cluster membership of the predicted cluster (i.e. the greater the cluster membership of the predicted cluster, the more the score of the performer is increased). Finally, the performer with the higher score is returned.

Clearly, the classifiers CL_i play a central role in the output of classifier *MC*. For each performer, CL_i is trained with data extracted from the performer's performance recordings. We have explored different classifier induction methods (described above) for obtaining each classifier CL_i. The whole procedure for training classifiers CL_i is as follows:

1. Collect all training recordings by all performers
2. Segment notes in the training recordings
3. For each segmented note *N,* compute its intra-note description *PN*
4. Using the intra-note description of all segmented notes, apply fuzzy k-means clustering (resulting in k clusters of notes, each cluster corresponding to a set of similar notes in terms of their intra-note description)
5. For each performer P_i,
    - Collect training recordings for that performer
    - For each segmented note *N* in the performer's recordings, compute *N*'s inter-note description *CN*
    - Build a classifier (e.g. a decision tree) using the inter-note features as attributes and its cluster (computed in step 4) as class.
6. Return the resulting classifier (e.g. the decision tree) CL_i for each performer P_i

The motivation for inducing the classifiers as described above is that we would like to devise a mechanism to capture which (perceptual) type of notes are played in a particular musical context by a performer. By clustering the notes of all the performers based on the notes' intra-note features, we intend to obtain a number of sets, each containing perceptually similar notes (e.g. notes with similar timbre). By building a decision tree based on the inter-note features of the notes of a performer, we intend to obtain a classifier which predicts what type of notes a performer performs in a particular musical context.

4.4 Evaluation

We evaluated the induced classifiers by performing the standard 10-fold cross validation in which 10% of the melody fragments is held out in turn as test data while the remaining 90% is used as training data. When performing the 10-fold cross validation, we leave out the same number of melody fragments per class. In order to avoid optimistic estimates of the classifier performance, we explicitly remove 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. Thus, the applied 10-fold cross validation procedure, in addition to holding out a test example from the training set, also removes repetitions of the example.

5. Case studies

In this section we present two case studies on identifying performers from their playing style, one in which the musicians performed the pieces by reading a score in a controlled studio environment, and another consisting of solo CD commercial performances. Note that the availability of the score in the first case study allows a complete analysis of the musical context of each performed note and enables us to establish a very complete mapping from this context to particular expressive transformations. However, in order to apply a unified methodology to both case studies (in the other case study the score of the performance is not available) we decided to discard the information provided by the score.

5.1 Controlled Studio Environment Performances

Training data

The training data used in this case study are monophonic recordings of four Jazz standards (*Body and Soul, Once I loved, Like Someone in Love* and *Up Jumped Spring*) performed by three different professional saxophonists in a controlled studio environment (the pieces were recorded in the Audiovisual Institute's recording studio at the Pompeu Fabra University, expressly for the experiment). The musicians were instructed to perform the selected pieces following a metronome and were asked not to introduce ornamentations. Each piece was performed at two different tempi (the tempi depended on the piece, e.g. for Body and Soul the tempi were 65 and 50 bpm). For each note in the training data, its intra-note features and inter-note features were computed. The performance tempo was added as an extra feature to the set of inter-note features.

Results

There were a total of 792 notes available for each performer. We segmented each of the performed pieces in phases and obtain a total of 120 short phrases and 32 long phrases for each performer. The length of the obtained phrases and long phrases ranged from 5 to 12 notes and 40 to 62 notes, respectively. The expected classification accuracy of the default classifier (one which chooses randomly one of the three performers) is 33% (measured in correctly classified instances percentage). In the short phrase case, the average accuracy and the accuracy obtained for the most successful trained classifier was 97.0% and 98.4%, respectively. In the long phrase case, the average accuracy and the accuracy obtained for the most successful trained classifier was 96.7% and 98.0%, respectively. The correctly classified instances percentage for each learning method is presented in Table 1. The results for short and long phrases seem to indicate that it is indeed feasible to train successful classifiers to identify performers from their playing style using the considered perceptual and contextual features. It must be noted that the performances in our training data were recorded in a controlled environment in which the gain level was constant for each performer. Some of the features (e.g. attack level) included in the perceptual description of the notes take advantage of this property and provide very useful information in the learning process. However, this recording requirement is not realistic in a general setting where we may obtain performances recorded under very different circumstances. In order to determine to which extent the results extend to a general setting, we consider commercial recordings from famous saxophonists in the next section.

	1-note	Short-phrase	Long-phrase
Decision Trees	37.4	95.1	95.8
Support Vector Machines	41.5	97.5	96.5
Artificial Neural Networks	39.8	97.5	95.6
k-Nearest Neighbor	31.2	97.5	96.5
Bagging (decision trees)	38.6	98.4	98.0
Boosting (decision trees)	39.4	95.1	96.2
Voting (decision trees, SVM, ANN, 1-NN)	42.7	97.5	97.2
Stacking (decision trees, SVM, ANN, 1-NN)	44.9	97.5	97.9

Table 1: Classification accuracy for the 1-note, short-phrase and long-phrase cases (in correctly classified instances percentage)

5.2 Solo Commercial Performances

Training data

The data used in this case study are monophonic audio commercial recordings of improvisations performed by four famous Jazz saxophonists: Billie Pierce (*Aria's Prance, Chelsea Bridge, In Your Own Sweet Way*), Joe Henderson (*Lush Life, Modinha*), Branford Marsalis (*St Thomas*) and Kenny Garrett (*Last Sax*). As opposed to the experiments reported in the previous section, here each musician performed a set of different pieces. Each of the recordings is analyzed as described before: the recordings are segmented into notes and for each note its intra-note and inter-note features are computed. In order to have a similar number of instances (notes and musical phrases) we selected a subset of the recordings.

Results

On average, there were a total of 820 notes available for each performer. We segmented each of the performed pieces in phases and obtain an average of 130 short phrases and 17 long phrases for each performer. The length of the short phrases and long phrases ranged from 4 to 10 notes and 30 to 60 notes, respectively. The expected classification accuracy of the default classifier (one which chooses randomly one of the four performers) is 25% (measured in correctly classified instances percentage). In the short phrase case, the average accuracy and the accuracy obtained for the most successful trained classifier was 71.9% and 74.7%, respectively. In the long phrase case, the average accuracy and the accuracy obtained for the most successful trained classifier was 71.1% and 74.9%, respectively. The correctly classified instances percentage for each learning method is presented in Table 2. The results for short and long phrases seem to indicate that the considered intra-note and inter-note features are indeed useful for training successful classifiers to identify performers from their playing style. It must be noted that the performances in our training data were recorded in a non controlled environment.

	1-note	Short-phrase	Long-phrase
Decision Trees	23.6	68.3	67.4
Support Vector Machines	27.6	74.7	74.9
Artificial Neural Networks	25.4	72.6	71.9
k-Nearest Neighbor	24.3	71.0	72.2
Bagging (decision trees)	26.4	70.4	66.3
Boosting (decision trees)	24.8	71.2	67.8
Voting (decision trees, SVM, ANN, 1-NN)	28.3	70.8	72.4
Stacking (decision trees, SVM, ANN, 1-NN)	29.0	76.3	76.3

Table 2: Classification accuracy for the 1-note, short-phrase and long-phrase cases (in correctly classified instances percentage)

Discussion

The difference between the results obtained in the case studies and the accuracy of a baseline classifiers, i.e. the classifier guessing at random, indicates that the intra-note and inter-note features presented contain sufficient information to identify the studied set of performers, and that the machine learning methods explored are capable of learning performance patterns that distinguish these performers. It is worth noting that every learning algorithm investigated (decision trees, SVM, ANN, k-NN and the reported ensemble methods) produced considerably better than random classification accuracies. This supports our statement about the feasibility of training successful classifiers for the case studies reported.

As mentioned before, the performances in the first case study were recorded in a controlled environment in which the gain level was constant for each performer. Some of the features (e.g. attack level) included in the perceptual description of the notes take advantage of this property and provide very useful information in the learning process. However, for the performances in the second case study we do not have information about the recording conditions (in fact the conditions surely were very different for different performers). This may explain the difference in accuracies between the first and second case studies.

We have selected three types of musical segment lengths: 1-note segments, short-phrase segments (4-12 notes), and long-phrase segment (30-62 notes). Evaluation using 1-note segments results in poor classification accuracies, while short-phrase segments and long-phrase segment evaluation results in accuracies well above the accuracy of a baseline classifier. Interestingly, there is no substantial difference in the accuracies for short-phrase and long-phrase segment evaluation which seems to indicate that in order to identify a particular performer it is sufficient to consider a short phrase segment of the piece, i.e. the identification accuracy does not increase substantially by considering a longer segment. The poor results of the 1-note evaluation may indicate that although intra-note features are very important, it is not sufficient to consider them in a one note basis. Just as a human Jazz expert would have problems identifying saxophonists form listening to one note audio files, the trained classifiers are not able to identify the performers reliably given this limited information. As soon as there are more notes involved together with the context in which they appear, the trained classifier (just as a Jazz expert) is able to identify the musician.

It is worth mentioning that ideally we would have liked to consider in all the experiments a data set containing the same set of pieces for each performer. However, in the second case study it was impossible to get hold of the same pieces by the different performers. This contrasts with other approaches (e.g. (Saunders, 2004)) in which the same set of performed pieces is available for the different performers.

One issue which is not clear from the reported case studies is what features are mostly responsible for the identification results. As mentioned before, we conjecture that a great part of the cues for performer identification in saxophone performances come from the timbre of the notes, i.e. the intra-note features. This is to say, while timing information is certainly important and is useful to identify a particular musician most of the information relevant for identifying a performer is the timbre characteristics of the performed notes. In order to investigate this hypothesis we have performed an additional experiment in which we have trained performer classifiers based only on the note-level timing and energy information in the performances. We have induced models predicting the timing (i.e. note duration) and energy (i.e. note mean energy) using the data from the controlled studio environment performances (first case study). For each performer *Pi* considered we have trained a model *Mi* predicting for a given note the pair *(Duration,Energy)* representing the duration transformation and mean energy variation for that note. As before, for each note in an input melody, we use this information to update score *Si* of performer *Pi.* After considering all notes in the melody, the performer with maximum score is the one identified by the system. In the short phrase case, the average accuracy and the accuracy obtained for the most successful trained classifier (obtained with the support vector machine algorithm) was 48% and 53%, respectively. In the long phrase case, the average accuracy and the accuracy obtained for the most successful trained classifier (obtained with the boosting algorithm) was 48% and 51%, respectively. These results seem to indicate that there is some performer-specific information in the duration-energy models but the models are certainly more accurate at identifying interpreters when intra-note information is included. Having said that, it is clear that these results depend on the set of performers under consideration, i.e. some performers may differ from each other mainly by their timbre while others may differ also in the way they apply dynamic and tempo transformations in their performances.

6. Conclusions

In this paper, we concentrated on the task of automatic identification of saxophone performers based on their playing style. We have applied sound analysis techniques to monophonic audio recordings in order to extract features such as pitch, timing, amplitude and timbre, characterising both the internal structure of notes and the musical context in which they appear. We explored and compared different machine learning techniques for this task. The main contribution of this work is the development of successful classifiers for the identification of performers not only on recordings obtained in a controlled environment, but on CD commercial recordings. The results obtained indicate that the extracted features contain sufficient information to identify the studied set of performers, and that the machine learning methods explored are capable of learning performance patterns that distinguish these performers. We plan to extend the set of intra-note descriptors with relevant descriptors such as vibrato and extend our approach to performance-based performer identification in polyphonic multi-instrument audio recordings.

References

(Van Baelen, 1996) Van Baelen, E. and De Raedt, L. (1996). Analysis and Prediction of Piano Performances Using Inductive Logic Programming. International Conference in Inductive Logic Programming, 55-71.

(Berstein, 1976) Bernstein, A. D., Cooper E. D. (1976). The piecewise-linear technique of electronic music synthesis. *Journal of Audio Engineering Society* (24) 6: 446-454.

(Bresin, 2002) Bresin, R. 2002. Articulation Rules for Automatic Music Performance. In Proceedings of the 2001 International Computer Music Conference. San Francisco, International Computer Music Association.

(Canazza, 1997) Canazza, S.; De Poli, G.; Roda, A.;and Vidolin, A. 1997 Analysis and Synthesis of Expressive Intention in a Clarinet Performance. In Proceedings of the 1997 International Computer Music Conference, 113-120. San Francisco, International Computer Music Association.

(Cano, 1998) Cano, P. 1998. “Fundamental Frequency Estimation in the SMS Analysis”. *Proceedings of the Digital AudioEffects Workshop (DAFx)*, Barcelona, 1998.

(Chauvin, 1995) Chauvin, Y. et al. (1995). Backpropagation: Theory, Architectures and Applications. Lawrence Erlbaum Assoc.

(Colmenauer, 1990) Colmenauer A. (1990). An Introduction to PROLOG-III. Communications of the ACM, 33(7).

(Cristiani, 2000) Cristianini N., Shawe-Taylor J. (2000). An Introduction to Support Vector Machines, Cambridge University Press

(Dannenberg, 1998) Dannenberg, R. B., and Derenyi, I. 1998. Combining Instrument and Performance Models for High-Quality Music Synthesis. Journal of New Music Research 27(3): 211-238.

(Dannenberg, 1998b) Dannenberg, R. D., Pellerin, H., Derenyi, “A study of trumpet envelopes” *Proceedings of the International Computer Music Conference (ICMC),* San Francisco, 1998

(Dovey, 1995) Dovey, M.J. (1995). Analysis of Rachmaninoff's Piano Performances Using Inductive Logic Programming. European Conference on Machine Learning, Springer-Verlag.

(Friberg, 2000) 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.

(Friberg, 2006) Friberg, A., Bresin, R., Sundberg, J. (2006). Overview of the KTH rule system for musical performance. Advances in Cognitive Psychology, Special Issue on Music Performance, 2(2-3), 145-161.

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

(Gabrielsson, 2003) Gabrielsson, A. (2003). Music Performance Research at the Millennium. Psychology of Music, Vol. 31, No. 3, 221-272 (2003)

(Gomez, 2003) Gómez, E., Klapuri, A., Meudic, B. (2003). Melody Description and Extraction in the Context of Music Content Processing. Journal of New Music Research. 32.

(Herrera, 1998) Herrera, P. Bonada, J., ”Vibrato Extraction and Parameterization in the SMS framework”. *Proceedings of COST G6 Conference on Digital Audio Effects (DAFx)*. Barcelona, 1998.

(Jenssen, 1999) Jenssen, K., “Envelope model of isolated musical sounds”. *Proceedings of COST G-6 Workshop on Digital Audio Effects (DAFx)*, Trondheim, 1999.

(Johnson, 1992) Johnson, M.L. (1992). An expert system for the articulation of Bach fugue melodies. In Readings in Computer-Generated Music, ed. D.L. Baggi, 41-51, IEEE Computer Society.

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

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

(Maestre, 2005) Maestre, E., and Gomez, E. 2005. “Automatic characterization of dynamics and articulation of monophonic expressive recordings” *In Proceedings of the 118th AES Convention*. Barcelona, Spain.

(Maher, 1994) 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.

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

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

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

(Narmour, 1991) Narmour, E. (1991). The Analysis and Cognition of Melodic Complexity: The Implication Realization Model. University of Chicago Press.

(Quinlan, 1993) Quinlan, J.R. (1993). C4.5: Programs for Machine Learning, San Francisco, Morgan Kaufmann.

(Ramirez, 2005) Ramirez, R. Hazan, A., and Maestre. E. 2005. “Intra-note features prediction model for jazz saxophone performance”. *In Proceedings of the 2005 International Computer Music Conference.* Barcelona, Spain.

(Ramirez, 2006) Ramirez, R. Hazan, A. (2006). A Tool for Generating and Explaining Expressive Music Performances of Monophonic Jazz Melodies, International Journal on Artificial Intelligence Tools, 15(4), pp. 673-691.

(Ramirez, 2006b) Ramirez, R., Hazan, A., Maestre, E., Serra, X. **“**A Data Mining Approach to Expressive Music Performance Modeling**”**, book chapter in *Multimedia Data mining and Knowledge Discovery*, Springer-Verlag

(Ramirez, 2008) Ramirez, R., Hazan, A., Maestre, E., Serra, X. (2008). A Genetic Rule-based Expressive Performance Model for Jazz Saxophone, Computer Music Journal, 32(1), pp.38-50.

(Repp, 1992) 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.

(Saunders, 2004) 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, 2004.

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

(Serra, 1990) 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.

(SMS) SMSTools: http://www.iua.upf.es/sms

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

(Tobudic, 2003)Tobudic A., Widmer G. (2003). Relational IBL in Music with a New Structural Similarity Measure, Proceedings of the International Conference on Inductive Logic Programming, Springer Verlag.

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

(Widmer, 2002) Widmer, G. (2002). Machine Discoveries: A Few Simple, Robust Local Expression Principles. Journal of New Music Research 31(1), 37-50.

(Widmer, 2001) Widmer, G. (2001). Discovering Strong Principles of Expressive Music Performance with the PLCG Rule Learning Strategy. Proceedings of the 12th European Conference on Machine Learning (ECML'01), Freiburg, Germany. Berlin: Springer Verlag.

(Widmer, 2004) Widmer, G., and Goebl, W. (2004). "Computational models of expressive music performance: The state of the art," Journal of New Music Research 33(3), 203-216.

(Witten, 1999) Witten, I.H. (1999). Data Mining, Practical Machine Learning Tools and Techniques with Java Implementation, Morgan Kaufmann Publishers.