Performance-based Interpreter Identification in Saxophone Audio Recordings

Rafael Ramirez, Esteban Maestre, Antonio Pertusa, Emilia Gomez, Xavier Serra

Abstract—We propose a novel approach to the task of identifying performers from their playing styles. We investigate how skilled musicians (Jazz saxophone players in particular) express and communicate their view of the musical and emotional content of musical pieces and how to use this information in order to automatically identify performers. We study deviations of parameters such as pitch, timing, amplitude and timbre both at an inter-note level and at an intra-note level. Our approach to performer identification consists of establishing a performer dependent mapping of inter-note features (essentially a 'score' whether or not the score physically exists) to a repertoire of inflections characterized by intra-note features. We present a successful performer identification case study.

***Index Terms*— Artificial Intelligence, Signal Processing, Music, Audio Recordings.**

I. Introduction

A key challenge in the area of music information, given the explosion of online music and the rapidly expanding digital music collections, is the development of efficient and reliable music search and retrieval systems. One of the main deficiencies of current music search and retrieval systems is the gap between the simplicity of the content descriptors that can be currently extracted automatically and the semantic richness in music information. Conventional information retrieval has been mainly based on text, and the approaches to textual information retrieval have been transferred into music information retrieval. However, music contents and text contents are of a very different nature which very often makes textual information retrieval unsatisfactory in a musical context. It has been widely recognized that music retrieval techniques should incorporate high-level music information.

In this paper we focus on the task of identifying famous performers from their playing style using high-level descriptors extracted from audio recordings. The identification of performers by using the expressive content in their performances raises particularly interesting questions but has nevertheless received relatively little attention in the past. Given the capabilities of current audio analysis systems, we believe expressive-content-based performer identification is a promising research topic in music information retrieval. This work is based on our previous work on expressive performance modeling (Ramirez, 2005) (Ramirez, 2006).

The data used in our investigations are audio recordings of real performances by famous Jazz saxophonists. The use of audio recordings, as opposed to MIDI recordings where data analysis is simplified, poses substantial difficulties for the extraction of music performance information. However, the obvious benefits of using real audio recordings widely compensate the extra effort required for the audio analysis. We use sound analysis techniques based on spectral models (Serra, 1990) for extracting high-level symbolic features from the recordings. The 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 high-level information such as note pitch, onset, duration, attack and loudness among other information. In particular, for characterizing structure in saxophone performances, we are interested in two types of features: *intra-note* features representing the internal structure of performed notes, and *inter-note* features representing information about the music context in which expressive events occur. We use the software SMSTools (SMS) which is an ideal tool for preprocessing the signal and providing a high-level description of the audio recordings. Once the relevant high-level information is extracted we apply machine learning techniques (Mitchell, 1997) to automatically discover regularities and expressive patterns for each performer. We use these regularities and patterns in order to identify a particular performer in a given audio recording. We discuss different machine learning techniques for detecting the performer's expressive patterns, as well as the perspectives of using sound analysis techniques on arbitrary polyphonic audio recordings.

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

Manuscript received July 21, 2006. This work was supported in by the Spanish Ministry of Education and Science under Grant TIN2006 (ProSeMus Project).

Rafael Ramirez, Esteban Maestre, Emilia Gomez and Xavier Serra are with the Music Technology Group, Pompeu Fabra University, Barcelona 08003, Spain. (corresponding author Rafael Ramirez, phone: 34935422165; fax: 34935422202; e-mail: rramirez@ iua.upf.edu). Antonio Pertusa is with the DLSI, Alicante University, Alicante Spain.

II. Background

Music performance plays a central role in our musical culture today. Concert attendance and recording sales often reflect people's preferences for particular performers. The manipulation of sound properties such as pitch, timing, amplitude and timbre by different performers is clearly distinguishable by the listeners. Expressive music performance studies the manipulation of these sound properties in an attempt to understand expression in performances. There has been much speculation as to *why* performances contain expression. Hypothesis include that musical expression communicates emotions (Justin, 2001) and that it clarifies musical structure (Kendall, 1990), i.e. the performer shapes the music according to her own intensions (Apel, 1972).

Understanding and formalizing expressive music performance is an extremely challenging problem which in the past has been studied from different perspectives, e.g. (Seashore, 1936), (Gabrielsson, 1999), (Bresin, 2002). 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. The majority of the research on expressive music performance has focused on the performance of musical material for which notation (i.e. a score) is available, thus providing unambiguous performance goals. Expressive performance studies have also been very much focused on (classical) piano performance in which pitch and timing measurements are simplified.

This paper describes a machine learning approach to investigate how skilled musicians (Jazz saxophone players in particular) express and communicate their view of the musical and emotional content of musical pieces and how to use this information in order to automatically distinguish among performers. 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 performance where timing and pitch measurements present a greater challenge compared to the measurements in piano performances (this is due to the fact that in piano performances certain expressive resources, e.g. vibrato and glissando, are absent).

Roughly, the basic idea of our approach to performer identification is to establish a performer-dependent mapping from inter-note features (essentially a 'score' whether or not the score physically exists) 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.

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. 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 where often the tempo of the performed pieces is not constant. The works focused on classical piano have focused on *global* tempo and loudness transformations while we are interested in both *intra-note* and *inter-note* level tempo and loudness transformations.

Widmer (Widmer, 2001) reported 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.

Tobudic et al (Tobudic, 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. Their learning algorithm recognizes similar phrases from the training set and applies their expressive patterns to a new piece.

Other inductive approaches to rule learning in music and musical analysis include (Dovey, 1995), (Baelen, 1996). In (Dovey, 1995), Dovey analyzes piano performances of Rachmaniloff pieces using inductive logic programming and extracts rules underlying them. In (Baelen, 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.

Nevertheless, the use of expressive performance models, either automatically induced or manually generated, for identifying musicians has received little attention in the past. This is mainly due to two factors: (a) the high complexity of the feature extraction process that is required to characterize expressive performance, and (b) the question of how to use the information provided by an expressive performance model for the task of performance-based performer identification. To the best of our knowledge, the only group working on performance-based automatic performer identification is the group led by Gerhard Widmer. Saunders et al (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.

III. Melodic description

In this section, we outline how we extract a description of a performed melody for monophonic recordings. We use this melodic representation to provide a inter-note and intra-note description of the performances and apply machine learning techniques to these extracted features. This is, our interest is to obtain for each performed note, a set of intra-note features and a set of inter-note features from the audio recording. The set of intra-note features includes descriptors such as 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. The set of inter-note features includes the relative pitch and duration of the neighboring notes (i.e. previous and following notes) as well as the musical structures to which the note belongs.

A. Extraction of inter-note features

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

As mentioned before, the main low-level descriptors used to characterize note-level expressive performance are instantaneous energy and fundamental frequency.

Energy computation. The energy descriptor is computed on the spectral domain, using the values of the amplitude spectrum at each analysis frame. In addition, energy is computed in different frequency bands as defined in (Klapuri, 1999), and these values are used by the algorithm for note segmentation.

Fundamental frequency estimation. For the estimation of the instantaneous fundamental frequency we use a harmonic matching model derived from the Two-Way Mismatch procedure (TWM) (Maher, 1994). For each fundamental frequency candidate, mismatches between the harmonics generated and the measured partials frequencies are averaged over a fixed subset of the available partials. A weighting scheme is used to make the procedure robust to the presence of noise or absence of certain partials in the spectral data. The solution presented in (Maher, 1994) employs two mismatch error calculations. The first one is based on the frequency difference between each partial in the measured sequence and its nearest neighbor in the predicted sequence. The second is based on the mismatch between each harmonic in the predicted sequence and its nearest partial neighbor in the measured sequence. This two-way mismatch helps to avoid octave errors by applying a penalty for partials that are present in the measured data but are not predicted, and also for partials whose presence is predicted but which do not actually appear in the measured sequence. The TWM mismatch procedure has also the benefit that the effect of any spurious components or partial missing from the measurement can be counteracted by the presence of uncorrupted partials in the same frame.

First, we perform a spectral analysis of all the windowed frames, as explained above. Secondly, the prominent spectral peaks of the spectrum are detected from the spectrum magnitude. These spectral peaks of the spectrum are defined as the local maxima of the spectrum which magnitude is greater than a threshold. The spectral peaks are compared to a harmonic series and a two-way mismatch (TWM) error is computed for each fundamental frequency candidates. The candidate with the minimum error is chosen to be the fundamental frequency estimate.

After a first test of this implementation, some improvements to the original algorithm where implemented to deal with some errors of the algorithm:

- Peak selection: a peak selection routine has been added in order to eliminate spectral peaks corresponding to noise. The peak selection is done according to a masking threshold around each of the maximum magnitude peaks. The form of the masking threshold depends on the peak amplitude, and uses three different slopes depending on the frequency distance to the peak frequency.
- Context awareness: we take into account previous values of the fundamental frequency estimation and instrument dependencies to obtain a more adapted result.
- Noise gate: a noise gate based on some low-level signal descriptor is applied to detect silences, so that the estimation is only performed in non-silent segments of the sound.

Note segmentation is performed using a set of frame descriptors, which are energy computation in different frequency bands and fundamental frequency. Energy onsets are first detected following a band-wise algorithm that uses some psycho-acoustical knowledge (Klapuri, 1999). In a second step, fundamental frequency transitions are also detected. Finally, both results are merged to find the note boundaries (onset and offset information).

Note descriptors. We compute note descriptors using the note boundaries and the low-level descriptors values. The low-level descriptors associated to a note segment are computed by averaging the frame values within this note segment. Pitch histograms have been used to compute the pitch note and the fundamental frequency that represents each note segment, as found in (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, by the following formula:

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

where $f_{ref} = 8.176$ (f_{ref} is a the reference frequency of the C_0). Then, we define histograms with bins of 100 *cents* and hop size of 5 *cents* and we compute the maximum of the histogram to identify the note pitch. Finally, we compute the frequency mean for all the points that belong to the histogram. The MIDI pitch is computed by quantization of this fundamental frequency mean over the frames within the note limits.

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

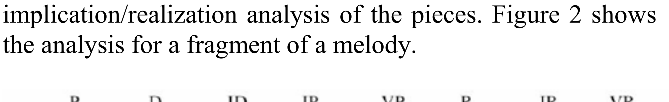

Fig. 1 Prototypical Narmour Structures

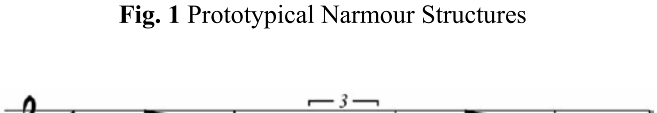

Fig. 2 Narmour analysis of *All of Me*

B. 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.

Intra-note segmentation. The proposed 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 3. 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 that 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.

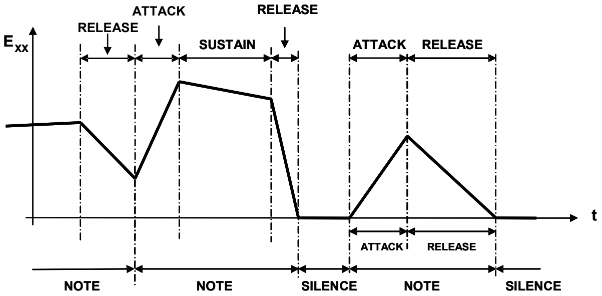

Fig. 3. Schematic view of the proposed energy envelope-based intra-note segmentation.

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 (2), 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.

$$e_m = \frac{1}{N}\sum_{k=1}^{N}\frac{\left|env(k) - env_m(k)\right|}{\overline{env}} \qquad (2)$$

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.

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 (3), where i and j denote frame positions. A minimum slope duration (measured in frames) Δfr 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.

$$\forall i, j \in F \text{ such as } i \leq j + \Delta fr, s_{i,j} = \frac{env_m(j) - env_m(i)}{j - i} \qquad (3)$$

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 (4) and (5) and depicted in figure 4.

$$s_M = s_{i_M, j_M} = \max(s_{i,j}) \ , \quad f_{AE} = j_M \quad (4)$$

$$s_m = s_{i_m, j_m} = \min(s_{i,j}) \ , \quad f_{RB} = i_m \quad (5)$$

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.

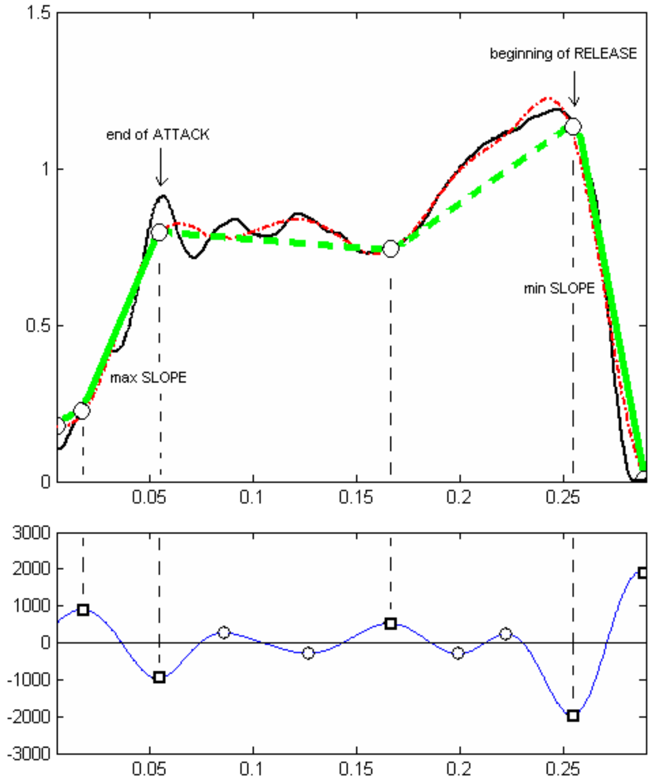

Fig. 4. Original and smoothed envelopes of a sax note for a value of e_{th}=0.05 (top figure, solid and dashed thin lines, respectively); selected characteristic points are denoted with a square within extremes of the second derivative of the smoothed envelope (bottom figure).

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 (6), where fft is the fast fourier transform of a frame, N is the size of the fast fourier tarnsform, 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.

$$SC = \frac{\sum_{k=1}^{N} k\left|fft(k)\right|}{\sum_{k=1}^{N}\left|fft(k)\right|} \qquad (6)$$

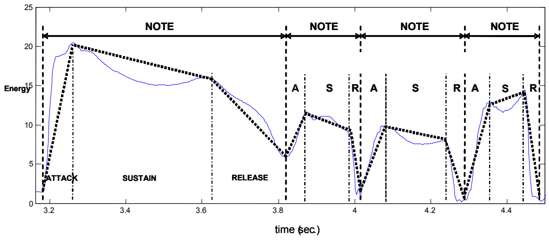

Fig. 5. Energy envelope and its linear approximation of a real excerpt with intra-note segment limits marked.

IV. Performance driven interpreter identification

In this section, we describe our approach to the problem of recognizing famous saxophonists from their playing style. In particular, we introduce the different note descriptors we use to characterize the internal and inter-note note properties (computed as described in the previous section), as well as the different algorithms we apply to identify performers from their playing style.

A. *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. The note's Narmour structures are computed by performing the musical analysis described in Section 3.1. Thus, each performed note is contextually characterized by the tuple

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

B. Algorithm

One of the first questions 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? In the case of Jazz saxophonists our hypothesis is that most of the cues for performer identification come from the timbre or 'quality' 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, Decision Trees (Quinlan, 1993), Support Vector Machines (SVM) (Cristiani, 2000), Artificial Neural Networks (ANN) (Chauvin, 1995), Lazy Methods, and Ensemble Methods.

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 $P_i$
    $Score_i = 0$
  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 $P_i$
      $Cluster^{i,k} = CL_i(CN_k)$
      $Score_i = Score_i + X_{Cluster^{i,k}}$
return $P_M$ such that $Score_M = \max(Score_1,\ldots,Score_n)$
```

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 $1,\ldots,q$, 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.

C. *Evaluation*

We evaluated the induced classifiers by performing standard test set validation in which a percentage of the melody fragments are held out in turn as test data while the remaining data is used as training data. When performing the 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 validation procedure, in addition to holding out a test example from the training set, also removes repetitions of the example.

V. Case Study

Important forms of performance in Western tonal music include performing music following a score, performing music by heart, performing improvised melodies, and playing by ear. With exception of the first form of performance, in the other forms of performance there is no notation (e.g. score) available. The task of identifying performers using the expressive information in their performances is only realistic if we consider performances for which we do not have the score the musician followed to produce the performance. Thus the question is: how to characterize the events in an expressive performance in order to capture their intra-note features and the musical context in which they appear? Our approach to this question is to study the intra-note features of an expressive performance by analyzing each note in a performance and building a performance alphabet of events, and by mapping the musical context in which the note appears to the symbols in the alphabet. In this way, we are able to describe a performance as a sequence of symbols in the performance alphabet and to characterize the musical context in which these symbols appear. A second question is: how to use this characterization in order to identify a musician in a new performance? Our approach to this question is to encode the new performance as a string of symbols in the performance alphabet and then to compare this string with the sequence of symbols each performer is expected to play.

In this section we present a case study on identifying performers from their playing style. We consider a set of monophonic recordings performed by reading a music score. Note that the availability of the score 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 obtain a unified methodology (in other case studies the score of the performance may not necessarily be available) we decided to discard the information provided by the score.

A. *Monophonic 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*) performed by three different professional saxophonists in a controlled studio environment. For each note in the training data, its inter-note and intra-note features were computed.

Results. We segmented each of the performed pieces in short and long phases for each performer. The length of the obtained phrases and long phrases ranged from 5 to 12 notes and 28 to 40 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 97.03% and 98.42%, respectively. In the long phrase case, the average accuracy and the accuracy obtained for the most successful trained classifier was 96.77% and 98.07%, respectively. The correctly classified instances percentage for each learning method is presented in Table 1. Clearly, the results for short and long phrases are statistically significant which indicates that it is indeed feasible to train successful classifiers to identify performers from their playing style using the considered intra-note and inter-note 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 intra-note description of the notes take advantage of this property and provide very useful information in the learning process. This recording requirement is not realistic in a general setting where we may obtain performances recorded under very different circumstances.

	1-note	Short-phrase	Long-phrase
Decision Trees	37.43	95.17	95.87
Support Vector Machines	41.50	97.50	96.58
Artificial Neural Networks	39.87	97.50	95.69
k-Nearest Neighbor	31.23	97.50	96.58
Bagging (decision trees)	38.67	98.42	98.07
Boosting (decision trees)	39.48	95.17	96.21
Voting (decision trees, SVM, ANN, 1-NN)	42.78	97.50	97.27
Stacking (decision trees, SVM, ANN, 1-NN)	44.92	97.50	97.93

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

B. *Discussion*

The difference between the results obtained in the case study and the accuracy of a baseline classifier, 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 significantly better than random classification accuracies. This supports our statement about the feasibility of training successful classifiers for the case study reported. However, note that this does not necessary imply that it is feasible to train classifiers for arbitrary performers.

We have selected three types of musical segment lengths: 1-note segments, short-phrase segments, and long-phrase segment. As expected, 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 sand 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.

VI. Conclusions

In this paper we focused on the task of identifying performers from their playing style using note descriptors extracted from audio recordings. In particular, we concentrated in identifying Jazz saxophonists and explored and compared different machine learning techniques for this task. We characterized performances by representing each note in the performance by a set of intra-note features corresponding to the internal structure of the note, and a set of inter-note features representing the context in which the note appears. We presented successful classifiers for a three-class classification task: identifying saxophonists in monophonic performances. The results obtained indicate 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. We are currently extending 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, 1975) Bernstein, A. D., Cooper E. D., “The piecewise-linear technique of electronic music synthesis”. *J. Audio Eng. Soc.* Vol. 24, No. 6, July/August 1976.

(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.

(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

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

(Friberg , 1998) Friberg, A.; Bresin, R.; Fryden, L.; and Sunberg, J. 1998. Musical Punctuation on the Microlevel: Automatic Identification and Performance of Small Melodic Units. Journal of New Music Research 27(3): 217-292.

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

(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.

(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.

(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, 2006) 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, 2005) Ramirez, R. and Hazan, A. (2005). Modeling Expressive Music Performance in Jazz, Proceedings of the 18th Florida Artificial Intelligence *Research Society Conference (FLAIRS 2005), Clearwater Beach, Florida.*

(Ramirez, 2005b) Ramirez, R., Hazan, A. (2005). A Learning Scheme for Generating Expressive Music Performances of Jazz Standards, Proceedings International Joint Conference on Artificial Intelligence, Edinburgh.

(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 , 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, 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.

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

Rafael Ramirez is an Assistant Professor at the Technology Department of the Pompeu Fabra University in Barcelona. He received a BSc in Mathematics from the National University of Mexico, an MSc in Artificial Intelligence from The University of Bristol, UK, and a PhD in Computer Science also from the University of Bristol. Prior to joining the Pompeu Fabra University he was a lecturer at the Department of Computer Science in the National University of Singapore, Singapore, and researcher at the National Research Institute in Computer Science (INRIA), France. His research interests include artificial intelligence, music information retrieval, declarative languages and music perception and cognition.

Esteban Maestre received his MA in Electrical Engineering from the Universitat Politècnica de Catalunya in 2003. After a research internship at Philips Research Laboratories Aachen, he worked as a teaching assistant in the Electronics Department at the Universitat Politècnica de Catalunya until 2004. Then, he joined the Music Technology Group at Universitat Pompeu Fabra as a research assistant, working also as a teaching assistant in the Technology Department, where he is currently a PhD candidate in Computer Science and Digital Communication. His research interests span from expressive audio analysis and synthesis to musical gestures coding and rendering.

Antonio Pertusa received the B.Sc. degrees in computer science and in computer systems (both in 2001), and the M.Sc. degree in computer science from the University of Alicante, Spain in 2003. After finishing his M.Sc., he joined the Department of Software and Computing Systems (Departamento de Lenguajes y Sistemas Informáticos) at the University of Alicante where he is currently
an assistant lecturer. He belongs to the Computer Music Lab, which is part of the Pattern Recognition and Artificial Intelligence Group at this University, and he is member of the Spanish Association of Pattern Recognition and Image Analysis. His research interests include machine learning, music information retrieval, signal processing and music perception and modeling.

Emilia Gómez is a researcher at the Music Technology Group of the *Pompeu Fabra University* (UPF). She graduated as a Telecommunication Engineer specialized in Signal Processing at the *Universidad de Sevilla*. Then, she received a DEA in Acoustics, Signal Processing and Computer Science applied to Music (ATIAM) at the IRCAM, Paris. She recently completed her PhD in *Computer Science and Digital Communication* at the UPF, on the topic of *tonal description of music audio signals*. During her doctoral studies, she was a visiting researcher at the Signal and Image Processing (TSI) group of the *École National Supérieure de Télécommunications* (ENST), Paris and at the Music Acoustics Group (TMH) of the *Stockholm Institute of Technology, KTH*.

Xavier Serra is Associate Professor at the Department of Technology of the Pompeu Fabra University (UPF) in Barcelona, Spain. He received the Ph.D. degree from Stanford University in 1989 with a thesis focused on spectral modeling of musical signals. Currently he is the director of the Music Technology Group of the UPF, a group dedicated to audio content analysis, description and synthesis.