Audio Engineering Society

Convention Paper

Presented at the 118th Convention
2005 May 28–31 Barcelona, Spain

Automatic Characterization of Dynamics and Articulation of Expressive Monophonic Recordings

Esteban Maestre, Emilia Gómez

Music Technology Group, Institut Universitari de l'Audiovisual, Universitat Pompeu Fabra
Ocata 1, 08003 Barcelona, Spain
http://www.iua.upf.edu/mtg
{emaestre,emilia.gomez}@iua.upf.es

ABSTRACT

We describe a method to automatically extract a set of features from the audio signal that are related to musical expressivity, more concretely to dynamics and articulation. We define a description scheme based on intra-note segmentation into attack, sustain, release and transition segments, and a subsequent amplitude and pitch contour characterization. Then, we present a series of algorithms to automatically perform intra-note segmentation and extract some features related to expressivity. We evaluate the performance of the methods for intra-note segmentation and feature extraction over a saxophone database of jazz standards and other recordings presenting expressive resources. Finally, we propose some future work and applications.

1. INTRODUCTION

Musicians do not play exactly what is written in the score. This fact makes music performance become an interesting topic of study. In its simplest sense, the term 'expression' is applied to those elements of a musical performance that depend on personal response and that vary between different interpretations or styles [1]. Performers enrich the sound using some expressive resources as for instance timing deviations (e.g. accelerando, ritardando), dynamics (e.g. *crescendo*, *decrescendo*), modulations (e.g. *vibrato*, *tremolo*) and articulations (e.g. *legato*, *staccato*). When studying music expressivity, it becomes necessary to measure those parameters in order to characterize the performance, i.e. how a certain piece has been played by a particular performer. Hence, many different parameters related to the audio signal are important descriptors when trying to analyze, transform and synthesize audio signals following concrete expressive patterns, as well as when trying to understand and model the style of a particular performer.

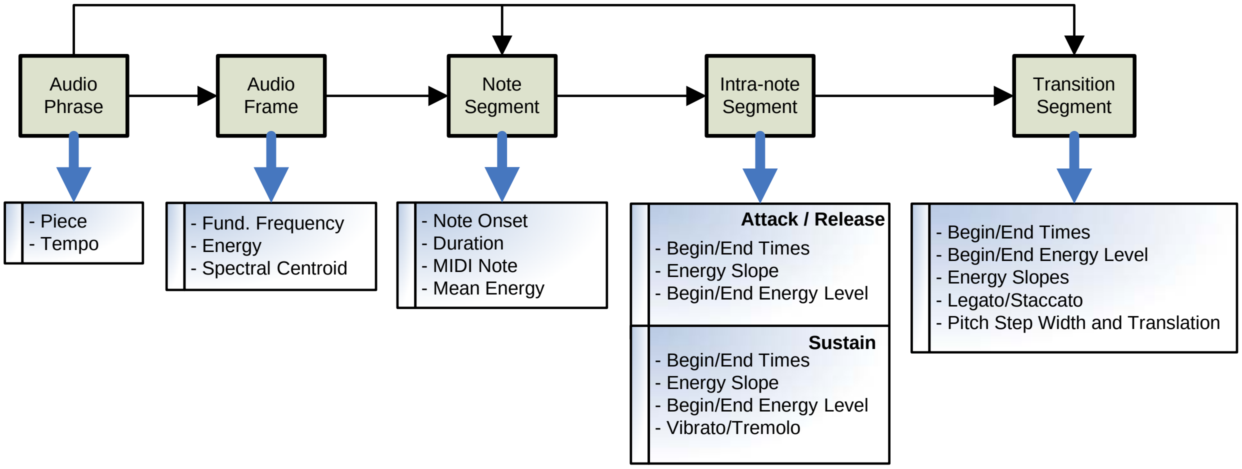


Figure 1. Description scheme.

This study has been developed in the context of a system intended to learn and transform expressive patterns from saxophone recordings of jazz standards [9]. Possible applications of this system include performer and performance understanding and modeling, as well as expressive audio transformations and synthesis [10]. In order to analyze the expressivity of an audio phrase, we first define a structured set of features related to expressivity (i.e. a description scheme) and then we find methods to extract these features in an automatic way. Previous work in [6] was devoted to extract a melodic description intended to analyze mainly timing deviations and insertions/ deletions of notes. This melodic description codes information related to the exact time location of each of the notes of the melody. In this work, we extend this melodic description to represent two important factors for music performance: dynamics and articulation. Other works in [4], [5] studied instrument specific temporal and spectral description for expressive synthesis.

Dynamics is usually defined as the intensity or volume with which notes are played. Articulation refers primarily to the degree to which a performer detaches individual notes from one another in practice (e.g. *staccato* and *legato*) [1]. In this study we present a method to automatically analyze dynamics and articulation from monophonic audio recordings of solo phrases, based in studying the energy envelope and fundamental frequency contour of the audio signal. The paper is structured as follows. First, we propose a scheme for expressivity description, focusing on dynamics and articulation. Then, the algorithms for intra-note segmentation and feature extraction are outlined. Section 4 includes a case study for which we evaluate the intra-note segmentation and feature extraction algorithms and present some experiments. Finally, some conclusions and future work are commented.

2. DESCRIPTION SCHEME

The proposed description scheme, which extends the scheme already presented in [6], is summarized in Figure 1. We define descriptors related to different temporal scales:

- Some features are defined as *instantaneous* or related to an analysis frame, such as energy, fundamental frequency [3] and spectral centroid.

- We also obtain intra-note/inter-note segment features, i.e. descriptors attached to a certain intra-note segment: attack, sustain and release (considering an ADSR model presented in [2]) or transition segment. We consider that the proposed features can set up a simple but concise dynamics and articulation model adapted to our application context. The procedure for the extraction of these features, which is the main focus of this work, is explained in Section 3.

- Note features or descriptors attached to a certain note are also extracted (already outlined in [6]).

- Finally, some global features are related to the whole performance, such as tempo or key.

3. FEATURE EXTRACTION

The basis of this study is the analysis of the energy envelope and fundamental frequency contour of the audio signal in order to extract features related to articulation and dynamics. The audio signal is divided into analysis frames, and some low-level frame descriptors are computed. Then, note segmentation is performed, extracting note features in order to obtain a melodic description. Finally, we carry out intra-note segmentation and characterize each intra-note and inter-note segments.

3.1. Melodic description

The first step consists of the extraction of a melodic description from the performance. This procedure, previous basis for this work, is explained in [6] and includes the extraction of low-level descriptors (energy, fundamental frequency and spectral centroid), the segmentation into notes and the computation of some note and global descriptors.

3.2. 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 [2]) 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 [8], 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, since our context is related to expressive patterns extraction by means of artificial intelligence techniques [6] and not a priori with complex synthesis, the model is simplified (in order to induce expression rules [9]), and 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.

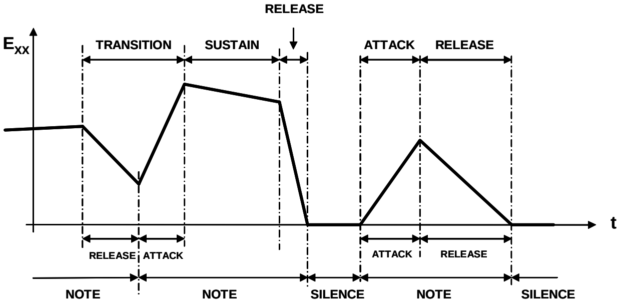


Figure 3. Schematic view of the proposed energy envelope-based intra-note segmentation.

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. 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 (1), 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.

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

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}. This threshold is empirically selected, as explained in next section. 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. In section 4, we evaluate the influence of both e_{th} and *n* on the algorithm performance. 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 (2), 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.

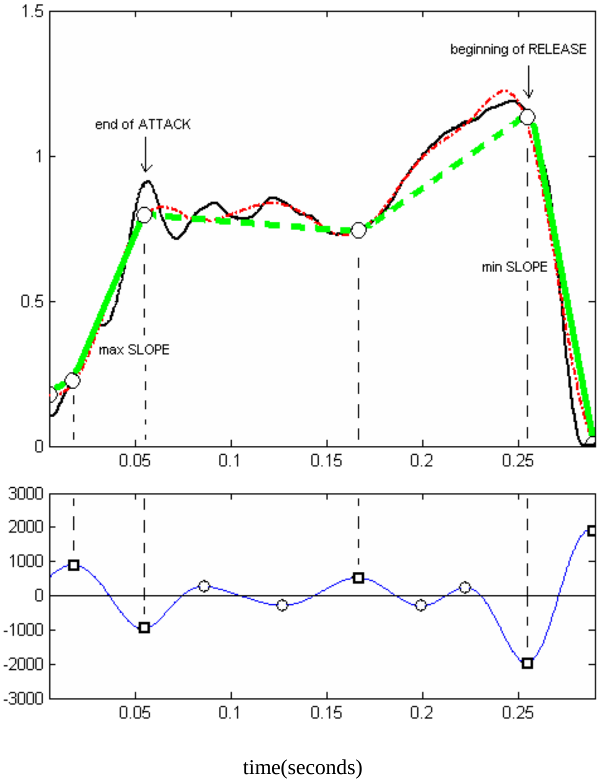


Figure 4. Original and smoothed envelopes of a real sax note extracted from the recordings for a value of e_{th}=0.05 (top figure, solid and dashed thin lines, respectively); selected characteristic points are denoted with a square within extremes of the second derivative of the smoothed envelope (bottom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

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 (2), 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.

$$\forall i, j \in F \text{ such as } i \leq j + \Delta fr, s_{i,j} = \frac{env_m(j) - env_m(i)}{j - i} \quad (2)$$

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

$$s_M = s_{i_M, j_M} = \max(s_{i,j}) \ , \quad f_{AE} = j_M \quad (3)$$

$$s_m = s_{i_m, j_m} = \min(s_{i,j}) \ , \quad f_{RB} = i_m \quad (4)$$

This algorithm can be schematically written as:

```
1.  Extract energy envelope
2.  Set initial $f_0=f_{0init}$
3.  Smooth the envelope with LP at $f_0$
4.  Compute $e_m$
5.  If $e_m>e_{th}$ then increase $f_0$ and go up to 3
6.  Obtain derivatives of the last smoothed envelope
7.  Get positions of modulus maxima of the 2nd derivative
8.  Select the n positions with highest 2nd derivative modulus
9.  Compute slopes between each pair of consecutive selected positions
10. Discard too short slopes
11. Get points defining both the maximum and minimum computed slopes
12. Assign attack-end to second point defining maximum slope
13. Assign release-begin to first point defining minimum slope
```

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 present sustain segment.

3.3. 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 sustain segment, we compute some vibrato and tremolo descriptors [7], consisting on start and end times, depth and rate. We tried several methods for estimating low-frequency modulations, obtaining better results by studying zero-crossings of the smoothed fundamental frequency and energy envelope derivatives.

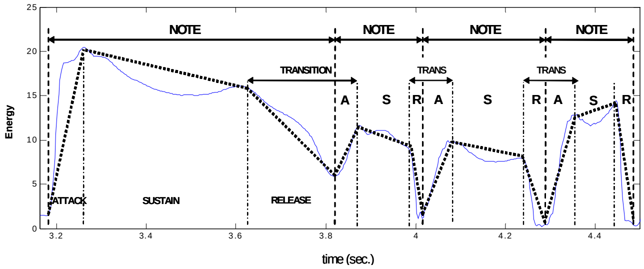


Figure 6. Energy envelope of a real excerpt with intra-note segment and transition limits marked.

3.4. Transition segment characterization

When there is no silence between two consecutive notes, we consider a note 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 during transitions are studied in order to extract descriptors related to articulation.

$$E_{TRANS\min} = \frac{t_c}{t_{end} - t_{init}} \qquad (5)$$

In addition to the characterization of release and attack segments of the involved notes, we measure the energy envelope minimum position t_c (see fig. 6) with respect to the transition duration as (5). This descriptor might be useful when reconstructing amplitude envelopes during transitions.

$$Legato_1 = \frac{A_1}{A_1 + A_2} = \frac{\sum_{t=t_{init}}^{t_{end}} L_t(t) - E_{XX}(t)}{\sum_{t=t_{init}}^{t_{end}} L_t(t)} \qquad (6)$$

In order to extract a descriptor indicating how sharply detached two notes were, we compute a *Legato* descriptor by means of two simple methods. For the first method, we join start and end points on the energy envelope contour by means of a line L_t that would represent the smoothest case of detachment. Then, we compute both the area below energy envelope and between energy envelope and the joining line L_t and define our legato descriptor as shown in (6).

$$Legato_2 = \frac{E_{T\min}}{\min(E_{RT}, E_{LT})} \qquad (7)$$

For the second method, we compute the ratio between the local minimum of the energy envelope within transition E_{Tmin} and the minimum of the values of energy envelope at transition boundaries E_{RT} and E_{LT} (7).

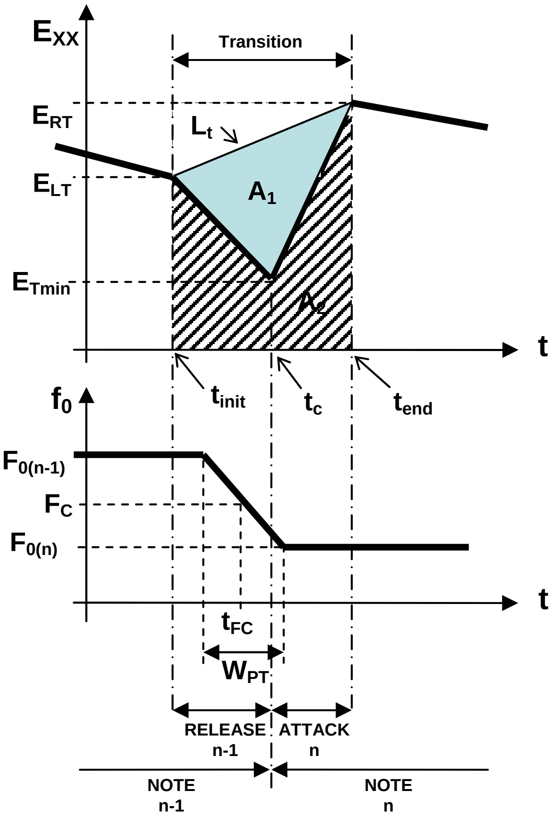


Figure 6. Schematic view of transition segment characterization

After observing fundamental frequency contours from the recordings, pitch steps are considered to be linear for

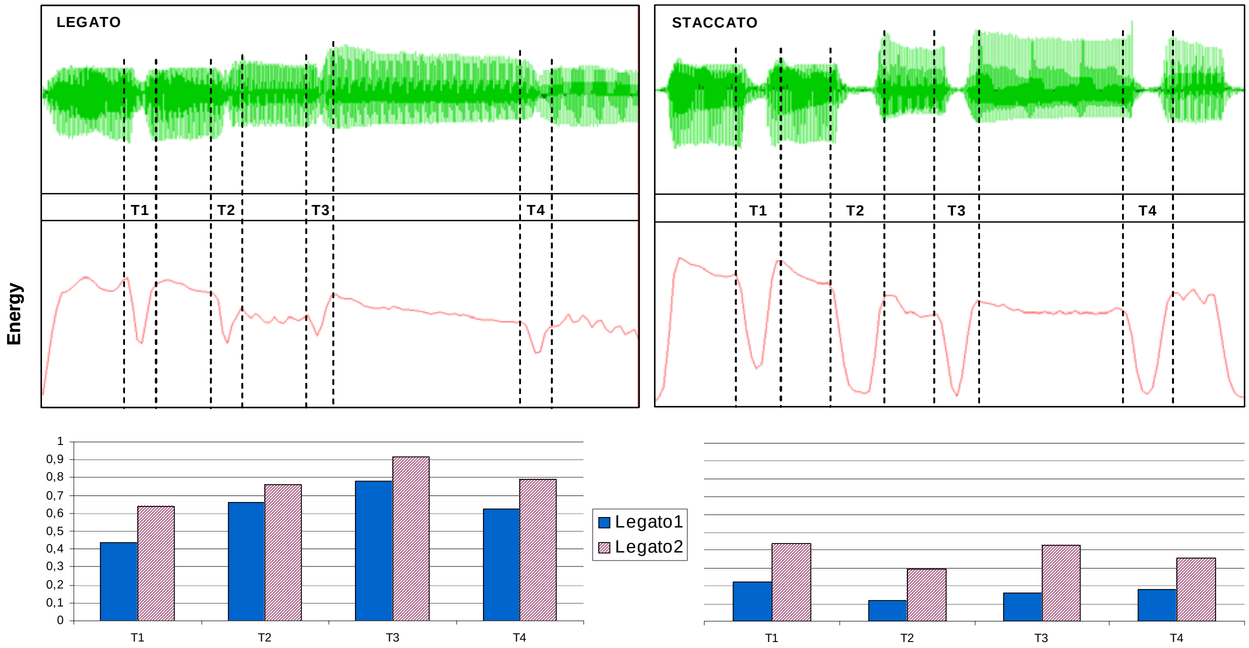


Figure 7. Schematic view computed legato descriptors for a saxophone excerpt performed by means of extreme legato/staccato articulations.

this study (see figure 6), being characterized by measuring width and translation with respect to the position of the energy envelope minimum and the transition length. Pitch step center time t_{FC} and width W_{PT} (see figure 6) are measured by finding the boundaries of pitch steps studying pitch derivative along the transition. Note that pitch step width W_{PT} is suitable for enrich a legato descriptor in terms of fundamental frequency description.

4. CASE STUDY

In this section, we show some evaluation results of the algorithm used for intra-note segmentation, and of both intra-note and transition segments characterization. For this study we have used an audio database consisting on 5 saxophone jazz standards played by a professional performer at 11 different tempi around the nominal one. Other saxophone recordings presenting expressive resources have been also used.

4.1. Intra-note segmentation

The results of the intra-note segmentation algorithm are compared to manual annotations, in order to evaluate its performance. This has been carried out with a set of 100 notes from our jazz performance recordings, where both the end of attack and the beginning of release have been marked. We compare the duration of the attack and release segments with the data extracted from the automatic segmentation for different values of intra-note segmentation algorithm parameters e_{th} and n (section 3).

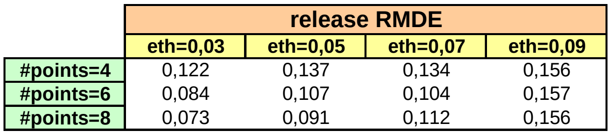
$$RMDE = \frac{1}{N}\sum_{n=1}^{N}\left|d_{RELsm}(n) - d_{RELsa}(n)\right| \quad (8)$$

We compute a segment relative mean duration error ($RMDE$), by means of (8), where N is the total number of notes, d_{RELsm} is the manually labeled duration for the corresponding segment relative to the current note duration, and d_{RELsa} is the automatically extracted duration, also relative to corresponding note duration.

	attack RMDE			
	eth=0,03	eth=0,05	eth=0,07	eth=0,09
#points=4	0,079	0,079	0,051	0,056
#points=6	0,080	0,080	0,053	0,068
#points=8	0,084	0,084	0,054	0,068

Table 1. Values of attack RMDE for different values of e_{th} and n used in intra-note segmentation algorithm.

The results obtained are shown in table 1 and table 2 for the attack and release segments respectively. They reveal reliable algorithm performance, and help in the selection of some parameters.



	release RMDE			
	eth=0,03	eth=0,05	eth=0,07	eth=0,09
#points=4	0,122	0,137	0,134	0,156
#points=6	0,084	0,107	0,104	0,157
#points=8	0,073	0,091	0,112	0,156

Table 2. Values of release RMDE for different values of e_{th} and n used in intra-note segmentation algorithm.

The increase of error for the case of release segment might be due to the fact that the algorithm sometimes assigns the whole sustain segment to be part of the release and vice versa (failing in detecting the presence of sustain segment leads to a high error), while for the attack segment there is less ambiguity.

4.2. Intra-note and transition segments characterization

Once intra-note segmentation gives accurate results, the validity of the feature extraction method is tested Regarding dynamics, we measured the error between the real energy envelope and the one obtained by approximating each intra-note segment by means of a linear segment (see figure 6). This error, that we call the approximation error (e_{APPR}), is computed as (9), where N is the number of frames of the current note, E_{REAL} is the real energy envelope of the note and E_{APPR} is the approximated energy envelope.

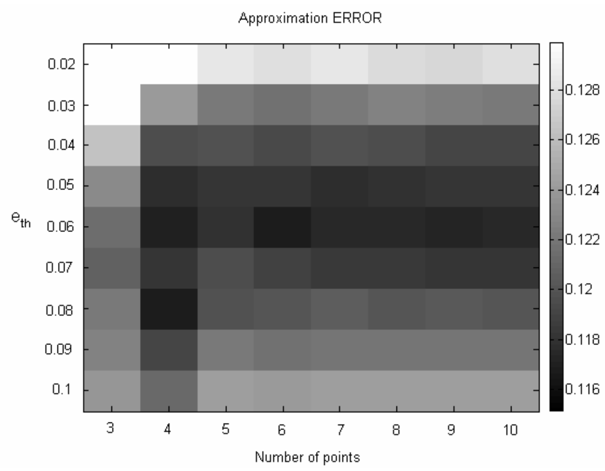


Figure 9. Mean approximation error for different values of e_{th} and n used in intra-note segmentation algorithm.

We have computed the average value of the approximation error for 1000 notes, using different values of the intra-note segmentation algorithm parameters e_{th} and n (section 3). As it can be observed in figure 9, the error threshold helps much more in finding a minimum, while the influence of the number of points is not so clear. Therefore, it is easy to find a compromise between the approximation error and the validity of the intra-note segmentation (section 4.1).

$$e_{APPR}=\frac{1}{N}\sum_{k=1}^{N}\frac{\left|E_{REAL}(k)-E_{APPR}(k)\right|}{\overline{E_{REAL}}} \qquad (9)$$

In terms of articulation, we have measured the relevance of the legato descriptors for the discrimination between two different expressive intentions. For doing so, a skilled saxophonist has been asked to play some phrases or excerpts articulating notes in both the extremes of *legato* and *staccato*, obtaining a high correlation between the computed descriptors and the expressive intentions, as it is shown in figure 7.

5. CONCLUSION AND FUTURE WORK

We have presented a description scheme for expressive performance of monophonic audio recordings. We have outlined a method to extract a set of expressive descriptors, in terms of energy and fundamental frequency, from saxophone expressive recordings. We have evaluated some of our results, for which our methods have shown to provide good reliability.

For intra-note and transition segments, in addition to analyzing other spectral descriptors, parametric curve models are to be extracted instead of linear segments, being very useful depending on the application and the accuracy and/or complexity of the model to be obtained. In terms of fundamental frequency contour description, a deeper study is still necessary in order to obtain reliable results. Some other instruments should be tried, and we surely have to modify and improve our methods for being as versatile and accurate as possible.

Although it is still in the analysis stage, we intend to build a performance model, in order to be able to reproduce the performance naturalness and/or expression, or even to study and identify different intentions or styles. Other applications may also include expressive MIDI mapping for existing synthesizers, as well as the extraction of key information for the development of new synthesizers.

6. ACKNOWLEDGEMENTS

This research has been partially funded by the Spanish TIC2003-07776-C02-02 project PROMUSIC, the EU-FP6-IST-507142 project SIMAC and the EU-FP6-eContent project HARMOS. The authors would like to thank Oscar Mayor and Pedro J. Ponce for their disinterested help in some saxophone recordings, and to Jordi Bonada and Alex Loscos for their useful advices.

7. REFERENCES

[1] "Grove Music Online" Online Publication, Oxford Univ. Press, 2004. http://www.grovemusic.com.

[2] 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.

[3] Cano, P. 1998. "Fundamental Frequency Estimation in the SMS Analysis". *Proceedings of the Digital AudioEffects Workshop (DAFx)*, Barcelona, 1998.

[4] Dannenberg, R. D., Pellerin, H., Derenyi, "A study of trumpet envelopes" *Proceedings of the International Computer Music Conference (ICMC),* San Francisco, 1998.

[5] Dubnov, S., Rodet, X. "Study of spectro-temporal parameters in musical performance" *Journal of New Music Research*, 2003.

[6] Gómez, E., Grachten, M., Amatriain, X., Arcos, J., "Melodic characterization of monophonic recordings for expressive tempo transformations", *Proceedings of Stockholm Music Acoustics Conference (SMAC),* Stockholm, 2003.

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

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

[9] Ramírez, R., Hazan, A., Gómez, E., Maestre, E., "Understanding expressive transformations in saxophone jazz standards using inductive machine learning". *Proceedings of Sound and Music Conference (SMC)*, Paris, 2004.

[10] Serra X., Bonada J.: "Sound Transformations based on the SMS High Level Attributes", *Proceedings of the COST G-6 Workshop on Digital Audio Effects (DAFx)*, Barcelona, 1998.