MELODIC CHARACTERIZATION OF MONOPHONIC RECORDINGS FOR EXPRESSIVE TEMPO TRANSFORMATIONS

Emilia Gómez[1], *Maarten Grachten*[2], *Xavier Amatriain*[1], *Josep Lluís Arcos*[2]

[1] Music Technology Group - Universitat Pompeu Fabra
Ocata, 1. 08003 Barcelona SPAIN
http://www.iua.upf.es/mtg
{emilia.gomez,xavier.amatriain}@iua.upf.es
[2] IIIA-CSIC-Artificial Intelligence Research Institute
CSIC - Spanish Council for Scientific Research
Campus UAB. 08193 Bellaterra, Barcelona SPAIN
http://www.iiia.csic.es
{maarten,arcos}@iiia.csic.es

ABSTRACT

The work described in this paper aims at developing a system that could perform expressive tempo transformations in monophonic instrument phrases. We have first developed a melodic description subsystem that extracts a set of acoustic features from monophonic recordings. This set of features is structured and stored following a description scheme that is derived from the current MPEG-7 standard. These performance descriptions are then compared with their corresponding scores, using *edit distance* techniques, for automatically annotating the expressive transformations performed by the musician. Then, these annotated performance descriptions are incorporated in a case-based reasoning (CBR) subsystem in order to build an expressive tempo transformations case base. The transformation subsystem will use this CBR system to perform tempo transformations in an expressive manner. Saxophone performances of jazz standards played by a professional performer have been recorded for this study.

In this paper, we first describe which are the melodic features that have been extracted and how they are structured and stored. Then, we explain the analysis methods that have been implemented to extract this set of features from audio signals and how they are processed by the CBR subsystem.

1. INTRODUCTION

As shown by Desain and Honing [1], the way a performer changes tempo in an expressive manner is not equivalent to performing a time-scale modification of the piece recorded in its nominal tempo. This is the fact that we address in this work. We want to study the deviations and changes that the performer introduces when changing the tempo of a piece. The study of these variations will be used as guidelines to perform tempo transformations in a expressive manner in the context of a content-based transformation system. Figure 1 represents the basic block diagram of the system. The audio processing part of the system consists of two different subsystems: the melodic description subsystem, that it is explained in section 2, and the transformation block, that is based in Spectral Modeling Analysis and Synthesis. The CBR part of the system basically consists of two subsystems: first there is the pre-processing part that consists of performance annotation, and musical processing (see subsection 3.1). In this part the cases are constructed. The second subsystem is the core reasoning procedure. It is explained in more detail insections 3.2, 3.3, and 3.4.

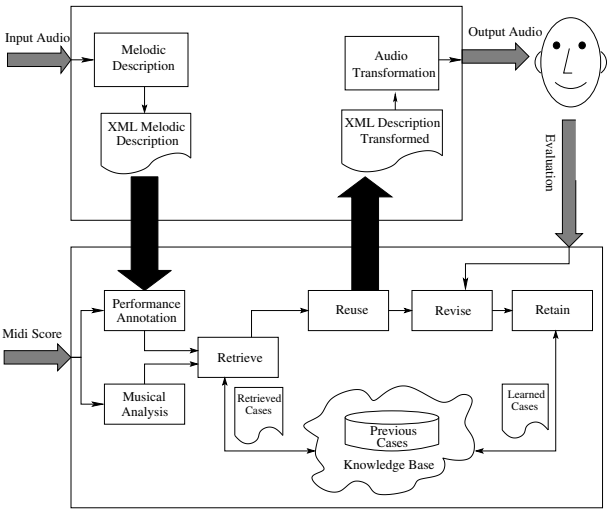

Figure 1: System architecture

2. THE MELODIC DESCRIPTION SUBSYSTEM

The melodic description block extracts melodic descriptions from monophonic recordings and stores them in a structured format. In this section we explain how the acoustic features are computed from audio and stored.

2.1. Description scheme

We define three structural levels of description: the analysis frame, the note and the global level. According to these structural levels,

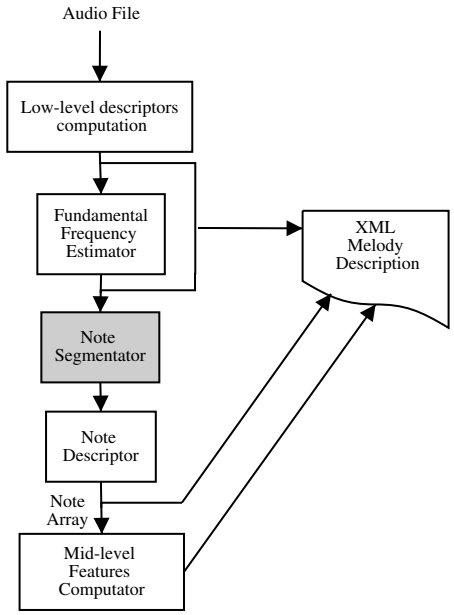

Figure 2: Block diagram of the melody descriptor

we define instantaneous descriptors, i.e. associated to an analysis frame, note descriptors attached to a note segment and global descriptors attached to the whole monophonic excerpt. All these descriptors are stored into a XML document. A detailed explanation about the description scheme and the MPEG-7 standard can be found in [2, 3].

2.2. Methods to extract the descriptors

Figure 2 represents the steps that are performed to obtain a melodic description from audio. First, the audio signal is divided into analysis frames, and the set of low-level descriptors are computed for each analysis frame. These low-level descriptors are used by the note segmentation algorithm, as well as in a preprocessing step of the fundamental frequency algorithm.

The fundamental frequency detector outputs an estimates for each analysis frame. Using these values and the low-level descriptors, the note segmentation block detects the note boundaries. Once the note boundaries are known, the note descriptors are computed from the low-level and the fundamental frequency values. Finally, note and low-level descriptors are combined to compute the global descriptors, associated to the whole audio segment.

2.2.1. Fundamental frequency estimation

For the estimation of the instantaneous fundamental frequency we use a harmonic matching model derived from the one proposed by Maher and Beauchamp [6], the Two-Way Mismatch procedure (TWM).

First of all, we perform a spectral analysis of a portion of sound, called analysis frame. 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. These spectral peaks are compared to a harmonic series and an 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 there 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-silences segments of the sound.

2.2.2. Note segmentation

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 [5]. In a second step, fundamental frequency transitions are also detected. Finally, both results are merged to find the note boundaries (onset and offset information).

2.2.3. 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 (as e.g. energy, centroid, spectral flatness, etc) 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 [7]. This is done to avoid taking into account mistaken frames in the fundamental frequency mean computation.

3. THE CASE BASED REASONING SUBSYSTEM

To apply musically meaningful transformations to the melodic descriptions obtained by the melodic analysis, it is of importance to know what musical material was actually performed and how the performance relates to this material. This information can be extracted from the score and the melodic descriptions of the performance. We constructed a CBR subsystem that uses this information to construct cases. These cases serve as ‘experience’ in the system. When a tempo transformation must be applied to a given musical performance, relevant cases are retrieved from the case base and using the retrieved cases a new performance is constructed at the desired tempo.

3.1. Construction of the cases

In the current system, a case consists of basically two kinds of information:

- A musical analysis of the score.
- An annotation of the performance that relates the performance to the score.

The musical analysis of the score is provided as a criterion for assessing similarities between scores and to split the phrases in smaller segments (see subsection 3.2). As an analysis-scheme, we use Narmour's Implication/Realization (I/R) model [9]. This is a model of melodic structure, based on principles akin to Gestalt Theory. An I/R analysis consists of a grouping of notes and categorizing these groups into a set of predefined categories. We have developed a parser for melodies that automatically generates I/R analyses. It implements most of the basic ideas from the I/R model.

The annotation of the performance is a reconstruction of how the musician actually performed the score. This annotation contains information such as which timing deviations were played, which score notes were not performed and which notes in the performance were not present in the score. Such an annotation can be conveniently obtained by mapping a melody description to the corresponding score. We use edit-distance techniques as described in [8] to generate mappings between the score and the melody description.

3.2. Retrieval of stored cases

The retrieval mechanism is organized in three phases:

3.2.1. Phrase retrieval

In a first phase the input melody (the score) is compared with the melodies of the case base using melodic similarity measures for retrieving only those case melodies really similar—For instance, given a slow ballad as input, we are not interested in comparing it with be-bop themes. To this end, similarities are computed between the input melody and each of the melodies in the case base. The similarities can be computed either using the note representations of the melodies, or using other representations such as melodic contour representations, or the musical models derived from the melodies. Combinations of these different measures can also be used to retrieve a subset of melodies. In [4], a comparison of various similarity measures is reported. It turns out that a similarity measure based on note level representations is not very discriminative when applied to melodies that are very different (i.e. all phrases that are not virtually identical, are assessed more or less equally dissimilar). Therefore, it seems more promising to use similarities between musical models or contour representations for retrieving similar melodies from the case base. The final output of this phase is a subset of a melodies of the case base close to the input melody. Only performances from these retrieved melodies will be taken into account in the following phases.

3.2.2. Motif retrieval

In a second phase, we try to find similar melodic fragments for segments of the input melody. The input melody is segmented based on the musical model that was constructed for the melody (including musical information such as the metrical strengths of the notes). In particular, I/R structures (or sequences of two or three structures) usually coincide with melodic motifs. For each of these segments, the most similar parts of the retrieved melodies can be selected (again using a similarity measure on either of the melodic representations). The result is that for each segment/motif of the input melody a set of melodic fragments is available, each fragment with one or more performance annotations.

3.2.3. Performance ranking

Finally, in the third phase the performances that were retrieved for each segment of the input melody are ranked using a similarity measure for the performance annotations. The idea behind this step is to use the input performance as a guide for how the input melody should be performed at the desired tempo. For illustration, say that an input performance P_{in} of a melodic segment M at tempo T_{in} was given and a new performance P_{out} of M at the desired tempo T_{out} must be generated. Suppose that there is a retrieved melodic fragment that has a performance P_1 at tempo T_i close to T_{in} and in addition, it has a performance P_2 at T_j close to the desired tempo T_{out}. Then, if performance P_{in} is similar to P_1, we assume that P_2 is a good basis for constructing performance P_{out} at tempo T_{out}. In this way, performances at tempos close to the desired tempo can be selected and ordered according to their expected relevance for constructing the solution.

In conclusion, the output of the retrieval step is an ordered collection of candidate annotations for each segment/motif in the input melody.

3.3. Reuse of retrieved cases

The reuse (or adaptation) mechanism is being implemented using *constructive adaptation* [10], a generative technique for reuse in CBR systems. The reuse mechanism deals with two kinds of criteria: local criteria and coherence criteria. Local criteria deal with the transformations to be performed to each note—i.e. how retrieved candidate annotations can be reused in each input note. Coherence criteria try to balance smoothness and hardness. Smoothness and hardness are basically contradictory: the first tends to iron out strong deviations with respect to the score, while the second tends to favor strong deviations. The resulting expressive performance is a compromise between the smoothness and hardness criteria, with the aim of keeping an overall balance pleasant to the ear.

3.4. Revising and Retaining cases

In the revise step, the user has the opportunity to either confirm the solution proposed by the system, discard it as being a bad solution, or manipulate it to correct local imperfections. When the result is satisfying to the user it is stored in the case base for solving future problems (the retain step).

4. AUDIO MATERIAL

For this experiment, we set up an audio database consisting of 5 jazz standards played at 11 different tempos around the nominal one, played by a professional musician. Most of the phrases where repeated to test consistency between performances. The jazz standards recorded where chosen to be of different moods: *Body and Soul*, *Once I Loved*, *Donna Lee*, *Like Someone In Love* and *Up Jumped Spring*.

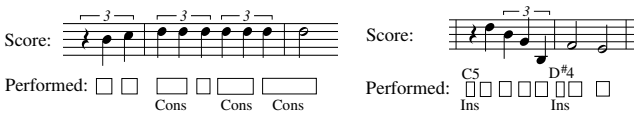

Figure 3: Two examples of the performer's interpretation of the score. *Cons* refers to a consolidation of notes, and *Ins* to a note insertion.

5. CASE-STUDY

We are currently in the process of analyzing the audio material, and constructing the CBR system. Nevertheless, in this section we will present some of the data we obtained by analyzing the audio recordings. First, we show two cases where the performer deviated from the score by inserting or consolidating (i.e. playing two or more notes as one) notes. On the left, figure 3 shows a fragment of *Once I Loved*, performed at 80 bpm. The blocks underneath the notes show what the performer played. It can be seen that the repetitions of the *D* pitch were not played literally, but instead some notes were played as one (this interpretation is preferred over the interpretation that some notes were *deleted*, since the rests caused by the omitted notes were filled up by an adjacent note). The example on the right shows a fragment of *Body and Soul*, also performed at 80 bpm. Here, the performer introduced two extra, ornamental notes. The inserted notes were of very short duration and the pitches of the notes were a semitone and whole tone below their successor notes, respectively.

Secondly, figure 4 shows the note onset deviations for performances of *Body and Soul* at different tempos. The y-axis shows the difference between performance onset and score onset, where negative values correspond to early performance onsets. It can be seen that the deviations are not constant under tempo changes. Nevertheless, there appear to be regularities. For instance, for the faster tempos, the beginning of the phrase was played with rather late timing, whereas for slower tempos the timing was neutral or slightly early.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have explained the set up of a system to automatically construct a knowledge base for expressive tempo transformations and use this knowledge base to perform expressive tempo changes to audio material. We also presented the results of the analysis and annotation of some audio recordings. This revealed that performances contained occasional deletions and insertions of notes, and that timing deviations tend to vary with tempo. Given the fact that the performances under consideration were rather neutrally played, this implies that a tempo transformation system for arbitrary jazz performances (which are often rather liberal interpretations of the score) cannot do without an analysis framework to interpret the performances. We intend to continue working on the analysis of the audio recordings to construct an initial case base. Moreover, the CBR system itself needs more elaboration and specification, in particular the Reuse step. Once the transformation system is fully operational, a user listening test could be set up in order to validate the transformation results.

Acknowledgments This work has been partially funded by the European IST Project CUIDADO (IST-1999-20194) and the Spanish TIC project TABASCO (TIC 2000-1094-C02).

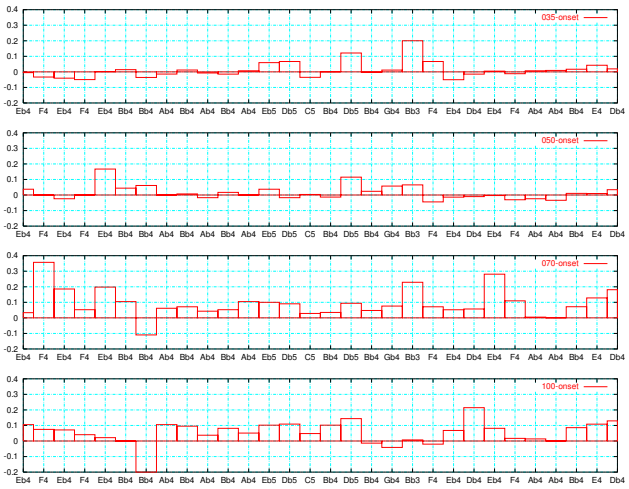

Figure 4: Onset deviations of *Body and Soul* for tempos 35, 50, 70 and 100 bpm.

7. REFERENCES

[1] Peter Desain and Henkjan Honing, *Tempo curves considered harmful. a critical review of the representation of timing in computer music*. Proceedings of the 1991 International Computer Music Conference, San Francisco, 1991.

[2] Emilia Gómez, Fabien Gouyon, Perfecto Herrera and Xavier Amatriain, *Using and enhancing the current MPEG-7 standard for a music content processing tool*, Proceedings of the 114th Audio Engineering Society Convention, Amsterdam, The Netherlands, March 2003.

[3] Emilia Gómez, Fabien Gouyon, Perfecto Herrera and Xavier Amatriain, *MPEG-7 for content-based music processing*, Proceedings of the 4th WIAMIS Special Session on Audio Segmentation and Digital Music, London UK, April 2003.

[4] Maarten Grachten, Josep Lluís Arcos, and Ramon López de Mántaras, *A comparison of different approaches to melodic similarity*. IInd International Conference on Music and Artificial Intelligence, 2002.

[5] Anssi Klapuri, *Sound Onset Detection by Applying Psychoacoustic Knowledge*, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 1999.

[6] Robert C. Maher and James W. Beauchamp, *Fundamental frequency estimation of musical signals using a two-way mismatch procedure*, Journal of the Acoustic Society of America, vol. 95 pp. 2254-2263, 1994.

[7] Roger J. McNab, Lliyd A. Smith and Ian H. Witten, *Signal Processing for Melody Transcription*,SIG working paper, vol. 95-22, 1996.

[8] Marcel Mongeau and David Sankoff, *Comparison of musical sequences*, Computers and the Humanities, 24:161–175, 1990.

[9] Eugene Narmour, *The Analysis and cognition of basic melodic structures : the implication-realization model*, University of Chicago Press, 1990.

[10] Enric Plaza and Josep Lluís Arcos, *Constructive adaptation*, In Susan Craw and Alun Preece, editors, Advances in Case-Based Reasoning, number 2416 in Lecture Notes in Artificial Intelligence, pages 306–320. Springer-Verlag, 2002.