MPEG-7 FOR CONTENT-BASED MUSIC PROCESSING*

Emilia GÓMEZ, Fabien GOUYON,
Perfecto HERRERA and Xavier AMATRIAIN
Music Technology Group, Universitat Pompeu Fabra, Barcelona, SPAIN
http://www.iua.upf.es/mtg
{*emilia.gomez,fabien.gouyon,perfecto.herrera,xavier.amatriain*}*@iua.upf.es*

The aim of this document is to present how the MPEG-7 standard has been used in a tool for content-based management, edition and transformation of audio signals: the *Sound Palette*. We discuss some MPEG-7 limitations regarding different musical layers, and some proposals for overcoming them are presented.

1. Introduction

The MPEG-7 standard provides content description utilities for audio and visual content, defining normative elements, as Descriptors (Ds), Description Schemes (DSs) and a Description Definition Language (DDL)[7]. Its Audio part[9] relies on two basic structures: the segment, which allows to define a temporal structure of the audio signal, and the scalable series, a type inherited by all the low-level descriptors (LLDs). It then distinguishes two classes of structures, the generic audio description framework (LLDs, the scalable series scheme and the silence segment) and the application-related tools (sound recognition, instrumental timbre description, spoken content description and melody description tools, as well as tools for audio matching).

The *Sound Palette* application, developed in the framework of the European project CUIDADO[a], is designed to be a an authoring tool for retrieving, editing, transforming and mixing isolated sound samples and phrases of reduced complexity (monophonic phrases, rhythm loops). The development of this application calls for a structured set of description schemes covering from signal-related descriptors to user-centered descriptors.

*This work has been supported by the european IST project CUIDADO and the spanish TIC project TABASCO.
[a]http://www.cuidado.mu

2. Melody description

The MPEG-7 standard distinguishes three types of attributes related to melody: the fundamental frequency LLD associated to a time point, melodic features attached to a *Note*, and the *Melody* DS and other support Ds related to the *AudioSegment* to which they are attached (see some description examples[1,4]).

The *Melody* DS[1,2,9] includes melody either as a melodic contour or a sequence of intervals plus note relative durations. The melodic contour uses a 5-step contour and it represents basic rhythm information by quantizing the duration of every note at the beat level. This contour has been found to be inadequate for some applications, as melodies of very different nature can be represented by similar contours[6]. For greater descriptive precision, the *Melody* DS supports an expanded descriptor set, where the precise pitch interval is kept and more accurate timing information is stored by encoding the relative duration of notes defined as the logarithm of the ratio between the differential onsets. Arranged around these core descriptors there are some optional support descriptors (lyrics, key, meter, and starting note).

This expanded description does not take into account silence parts that sometimes play an essential role for melodic perception. It is strongly tied to score representation and does not provide a direct link with LLDs. Regarding the *Note* representation, features like e.g. intensity, articulation or vibrato are lacking; they nevertheless would be necessary, for example, for expressivity characterization. Also, the *Note* is always defined as a part of the *noteArray* in the context of a *Melody*. Regarding the *key* and *scale* descriptors, MPEG-7 does not consider possible changes inside the segment (this is a general consideration for all the unary descriptors). Other scalar descriptors could also be added to the *Melody* DS according to the type of sounds and to application needs[6].

3. Rhythm description

MPEG-7 rhythmic elements are the *MeterType*, the *BeatType* and the note relative duration. Both the latter are embedded in the melody description (respectively in the contour and in the note). The *BeatType* represents to which beat "pertains" each note, i.e the quantized note positions with respect to the first note of the excerpt, expressed as integers, multiples of a timing reference, the *Beat* –which value in seconds is not given. The *MeterType* carries in its denominator a reference value for the expression of the beat series. This rhythmic representation has been proved useful

for query-by-humming applications. But let us wonder which could its limitations be for other applications.

First, the context is that of a monophonic melody, which seems to be a restriction. Then, the rhythm of a score, of a MIDI stream, or of an audio signal cannot be represented without loss of information. The time signature can be represented, but not the bar lines. The quantization w.r.t. the beat does permit to represent symbolic durations (e.g. quarter-note); nor does the note relative duration permit to account for exact note timing occurrences (which would serve e.g. for exploring expressivity deviations from the rhythmic structure). The speed of execution of a musical piece (the tempo) is absent from the standard. A proposal of tempo descriptor has been made to MPEG: the *AudioTempo*[8]. It is defined as a scalar value (number of BPM). It can be useful as a global descriptor of a piece of music, to account for its global pace. However, the evolution of the tempo is also a very important rhythmic feature, and representing it by means of the *AudioTempo* would entail a segmentation in audio segments whose only reason of being would be their tempo differences. This is questionable. Assuming constant tempo being relevant, another important feature is lacking: the phase of the beat. More important, improving the current standard by adding a single metrical level forgets the fundamental notion of hierarchy in the rhythmic structure of music. Finally, current MPEG-7 rhythmic Ds are extremely sensitive to the determination of the meter, which still remains a difficult task; no algorithms are suggested for its determination (informative extraction procedures are provided by MPEG for many other descriptors).

4. Instrument description

MPEG-7 provides Ds and DSs for timbre as a perceptual phenomenon, useful in the context of search by similarity in sound databases. These Ds, grouped together with other LLDs are: *HarmonicSpectralCentroid, HarmonicSpectralDeviation, HarmonicSpectralSpread, HarmonicSpectralVariation, SpectralCentroid, TemporalCentroid*, and *LogAttackTime*. They assume that segments can be allocated to generic "percussive" or "harmonic" classes, and after that identification, some specific weighting grants a retrieval by perceptual similarity.

Complementarily, some Ds and DSs permit to address verbal descriptions, allowing to perform categorical queries in databases or to build taxonomies of instruments. The *ClassificationScheme*[5,3] defines a scheme for

classifying a subject area with a set of terms organized into a hierarchy. A term is referenced in a description with the *TermUse* datatype. A term represents one well-defined concept in the domain covered by the classification scheme. It has an *identifier*, a *name*, and a *definition*. Terms can be put in relationship with a *TermRelation* descriptor, which represents a relation between two terms in a classification scheme. When terms are organized this way, they form a classification hierarchy.

Giving the user the option for defining instrument or sound taxonomies is an important feature, but populating a database with terms from them can be tedious unless would be provided an automatic linking mechanism using some mathematical model for computing the name of the class, provided some LLDs. This mechanism is restricted in the current version of MPEG-7 to the timbral description of the audio through a continuous Hidden Markov Model using a low-dimensional representation of the spectrum, the *SpectrumBasis*. This descriptor contains basis functions used to project spectrum descriptions into a low-dimensional and decorrelated representation. The basis functions are estimated through singular value decomposition, although other methods could be considered. This approach disregards other possibly valid options for sound modelling.

5. Using and enhancing MPEG-7

In its current version, MPEG-7 can be used for some applications (e.g. query-by-humming, timbral similarity search), however it still needs enhancements to cover a wider range of applications. In our context, we encountered a problem in the too general definition of the *AudioSegment*. It is our belief that defining specific temporal scopes of description (with different melodic, rhythmic and instrumental descriptors) would open the way to more accurate descriptions of musical excerpts. A way to deal with limitations could be to derive two types of segments: a *NoteSegment* (in the authors' opinion, a note has to be conceptually considered as a segment, not a descriptor) and a *MusicSegment* (representing a monophonic or polyphonic excerpt, being possibly decomposed into Note or Music segments) (see Figure 1). Note that there are some changes with respect to previous proposals[5], we derive from the abstract class *Segment* and not from *AudioSegment*. Another possible solution would be to define a hierarchy of segments by means of a *ClassificationScheme*[5]. Figure 2 and 3 show the DSs and Ds associated to both types of segments[5]. An extension of the *SoundModelType*, showed in Figure 4, is used for expanding the modelling

options and allowing other common ways of representing in a compact way the criteria for assigning a class to a given audio segment.

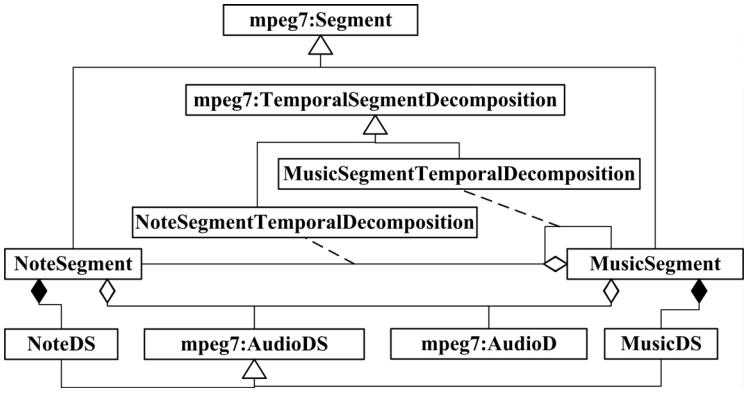

Figure 1. Segment definitions.

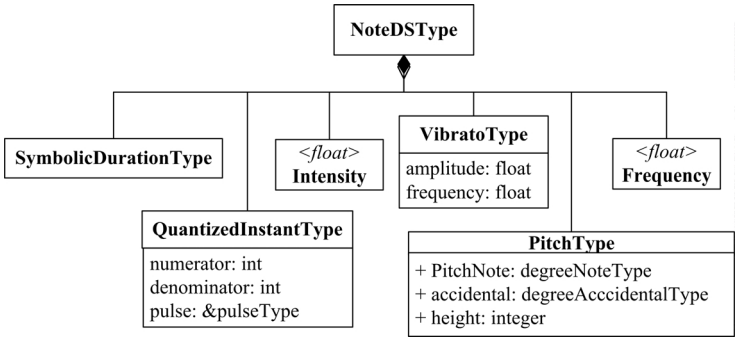

Figure 2. Note DS class diagram.

6. Conclusions

We intended to cope with the description requirements of the *SoundPalette* application. We have still left out issues of harmony or emotional load descriptions, as they do not seem to be priorities in our context. Extensions of the current standard have been proposed keeping in mind the need for compatibility; they should be considered as the beginning of an open discussion regarding what we consider as the current shortcomings of MPEG-7.

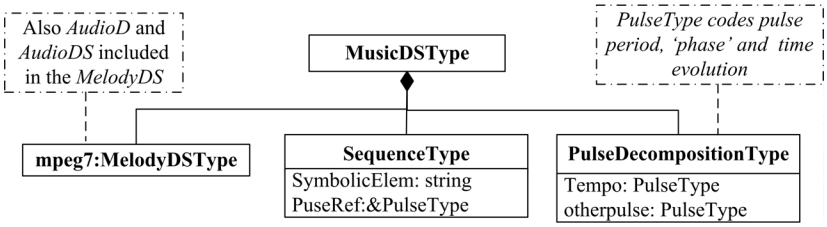

Figure 3. Music DS class diagram.

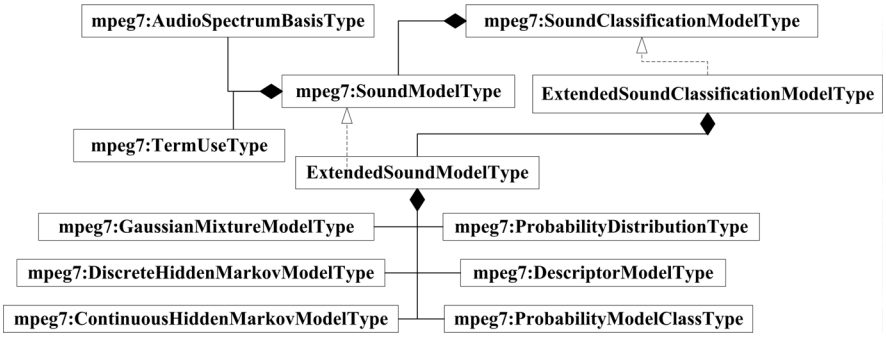

Figure 4. ExtendedSoundClassificationModel and ExtendedSoundModel definition.

References

1. *MPEG Working Documents.* http://www.cselt.it/mpeg/working_documents.htm.
2. *MPEG-7 Schema and description examples.* Final Draft International Standard (FDIS), http://pmedia.i2.ibm.com:8000/mpeg7/schema/, 2002.
3. M. A. Casey, *General sound classification and similarity in MPEG-7.* Organized Sound **6**, pp. 153-164, 2001.
4. E. Gómez, *Melody Description Scheme.* http://www.iua.upf.es/ egomez/mpeg7, last updated december 2002.
5. E. Gómez, F. Gouyon, P. Herrera and X. Amatriain, *Using and enhancing the current MPEG-7 for a music content processing tool*, 114th AES Convention, March 2003.
6. E. Gómez, A. Klapuri and B. Meudic, *Melody Description and Extraction in the Context of Music Content Processing*, Journal of New Music Research Vol. 32 .1, 2003.
7. W. Haas and H. Mayer, *MPEG and its Relevance for Content-based Multimedia Retrieval.* Journal of Universal Computer Science **7**, pp. 530-547, 2001.
8. J. Herre, M. Cremer, C. Uhle and J. Rohden, *Proposal for a core experiment on AudioTempo.* MPEG2001/8415 2002.
9. A. T. Lindsay and J. Herre, *MPEG-7 and MPEG-7 Audio - An Overview.* AES Journal **49**, pp. 589-594, 2001.