Audio Descriptors and Descriptor Schemes in the Context of MPEG-7

Perfecto Herrera, Xavier Serra
Audiovisual Institute – Pompeu Fabra University
Rambla 31, 08002 Barcelona, Spain
http://www.iua.upf.es
{pherrera, xserra}@iua.upf.es

Geoffroy Peeters
IRCAM
1 Place Igor-Stravinsky, 75004 Paris, France
http://www.ircam.fr
Geoffroy.Peeters@ircam.fr

[Published in the Proceedings of the ICMC99]

Abstract

Sound content description is one of the aims of the MPEG-7 initiative. Although MPEG-7 focuses on indexing and retrieval of audio, there are other sound content-based processing applications waiting to be developed once we have a robust set of descriptors and structures for putting them into relation and for expressing semantic concerns about sound. Spectral Modeling techniques provide a valuable framework for extracting and organizing sound content descriptions. All our descriptors can be considered low- or mid-level, for which we have devised automatic extraction procedures. In this paper we present the basic descriptors and introduce a scheme for organizing them (a so called "Descriptor Scheme"). We do not cover the high level description of music (musical forms and styles, roles of characters in a movie, etc.) which is also relevant in MPEG-7.

1. Introduction

MPEG-7 is an standardization initiative of the Motion Pictures Expert Group that, instead of focusing on coding like MPEG-1, MPEG-2 and MPEG-4, is meant to be an standardization of the way to describe multimedia content (MPEG-7 1999). The main application for MPEG-7 should be the content-based indexing and retrieval of audio, or of any other media. Content based audio descriptions can prove extremely fruitful in conventional tasks such as audio editing, music composition, sound effects selection, or video cueing, but also they can open new possibilities for live music mixing and DJing, sound signaturing for copyright protection, music commercial assessment and recommendation, agent based TV scheduling, etc.

Audio content extraction and managing can be achieved by different, sometimes overlapping, and anyway non-exclusive means: there are traditional signal processing techniques (Wold et al. 1996), computational auditory scene analysis techniques (Wyse and Smoliar 1998; Martin 1998), statistical techniques (Foote 1997) etc. There are also manual keyword labeling techniques and thesauri that can work synergistically with automatic techniques in order to help to describe and organize more thoroughly the content of audio material. The key issue here is that MPEG-7 will not standardize the way to obtain these descriptions nor how to use them, but only the descriptions and the way for structuring them.

Describing sound content involves the use of procedures, techniques, and data, that have been found and developed in different research areas (i.e. signal processing, music cognition, artificial intelligence, etc.), for solving related problems. Our work on sound analysis/synthesis with SMS (Serra and Bonada 1998) and other related techniques, result into sound measures useful as descriptors for the content of sound. As it has been showed in relevant literature (Wold et al. 1996; Zhang and Kuo 1998) combinations of them can be used as a basis for elaborating higher-level descriptions that approximate ordinary ways of referring to sounds in search, classification, and transformation tasks.

Complex audio content descriptions need a way for structuring and handling those descriptors besides the

additional information that could be derived from these new data organizations. We will discuss the characteristics of such a *descriptor scheme*, considering that sound descriptions are not only useful for high level searches (i.e. search for male voices, or for guitar solos), but also macro-structural and micro-structural (search for timbres with similar variations in the partial amplitudes along time, for example). As we will see, the description of an audio file with the help of spectral models can be done at multiple levels: we can describe it frame-by-frame, with so-called instantaneous or low level descriptors, but also it can be done at a higher scale, taking into consideration a temporal *segment* of a file or stream, or a spectral *region* along a segment. We believe that the efficient exploitation of such a kaleidoscopical representation of sound can give way to descriptions of sound content usable for solving many problems in multimedia applications.

2. Audio Descriptors

The descriptors that result from current spectral modeling approaches, such as instantaneous frequency, amplitude and phase of each partial or the instantaneous spectral characteristics of the residual signal, account for the microstructure of a sound. But, starting from them, there are also other useful instantaneous attributes that give a higher level abstraction of the sound characteristics (Serra and Bonada 1998). These attributes are easily calculated at each analysis frame from the output of, for example, the basic SMS analysis. Examples of these attributes are:

- Amplitude of sinusoidal component
- Amplitude of residual component
- Spectral shape of the sinusoidal component
- Spectral shape of the residual component
- Harmonic distortion
- Noisiness
- Spectral centroid
- Spectral tilt

Each descriptor has four different measures: instantaneous value (one frame), instantaneous variation (from frame to frame), mean value (in a temporal segment) and variance (in a temporal segment), thus characterizing different aspects of the sound. Other descriptors that we have considered are: Odd/even partials ratio, Tristimulus (Pollard and Jansson 1982), and Attack's harmonic coherence (Grey and Moorer 1977).

For a meaningful parameterization we have to segment the sound into fragments that are homogeneous in terms of certain sound attributes. Then we can identify segment descriptors that will give a summary of its content, and may allow to classify the segment into semantic categories corresponding to sound events or sound objects. One of the most obvious and general segmentation processes divides a musical melody into notes and silences and then each note into an attack, a steady state and a release segments. Global descriptors that can characterize attacks and releases refer to the statistical measures of each of the instantaneous descriptors, such as average fundamental frequency variation, average amplitude variation, or spectral centroid trajectory. In the steady state segments it is meaningful to characterize other global attributes such as time-varying rate and depth of vibrato and tremolo (Herrera and Bonada 1998). In summary, when we segment sound, we are using low-level content descriptors that can also help to characterize – when appropriately combined and interpreted- meaningful segments of sound. Contrasting with instantaneous measures, the segment description uses means, variances and derivatives in order to get the "global picture" along its duration.

As a final issue, it should be noted that the descriptors used in our spectral modeling environment allow a big degree of overlapping regarding its description power. This kind of overlap has also been positively considered in experimental systems specialized in audiovisual content analysis as the MOCA project (Pfeiffer et al. 1996). We believe that a standard for content description as MPEG-7 should accommodate this kind of redundancy, leaving to the front-end or content providers's proprietary software the effective use of the descriptors set.

3. Audio Descriptor Schemes

As a way of organizing our own descriptors and others that might be proposed by different researchers or users of an audio content search and classification software tool, we have devised what in MPEG-7 is called a "Description Scheme", that is, a structure that specifies the semantics and the relationships between Descriptors and other Description Schemes. The scheme started as an extension of SDIF (Sound Description Interchange Format) (Wright et al. 1998) an ongoing proposal for the storage of spectral analysis data, but it has evolved into a very different initiative.

The scheme we present here addresses the description of multichannel, multi-source sounds. It is intended to be a generic description scheme based on a hierarchical (and recursive in some places) structure, that can describe sound at multiple levels of detail (from the low level of a FFT Analysis Frame to the high level of a whole sound file), accommodating different kinds of dependencies and relationships among its components. Given its modularity and expandability it can be used to describe any kind of sound, and it can take into account different kinds of descriptors of sound. Descriptions can be extracted

and stored as needed and depending on the target application, and are not compulsory at all. This way, the scheme can contain descriptions at only one level, or at all available levels of description.

The scheme has two main levels: *syntactic* and *semantic*. We need a syntactic level in order to describe the sound file or stream in a superficial, temporal and structural way. But we also need a semantic level in order to assign semantic labels to the elements that deserve a description. The syntactic level is used to specify physical structures and the signal properties of the sound program and can be considered as a table of contents. The elements of the syntactic part of the scheme are *tracks*, *segments*, specific instances of segments like *frames*, and *regions*. On the other side, the semantic level is used to specify semantic features of a sound program in terms of *audio events* and *audio objects*, and can be viewed as a set of indexes. Another way of interpreting this twofold structure would be considering the syntactic level as the container for low-level descriptors and the semantic level as a container for mid and high level descriptors, although following this interpretation as it is could be a bit controversial under some circumstances.

3.1 Descriptor Schemes for the syntactic level

Track Description Scheme

A first substantial difference between Video media and Sound media is that sound can be expressed along several channels at the same time. Those channels may contain related information but also very distinct ones, so it is clear that we need a scheme to describe each Track separately while allowing links between Track descriptions.

A Track DS is then described by:

- **Time DS**, describes the beginning and the end of the track.
- **Placement DS**, describes the spatial placement of the track from the listener point of listening.
- **Media information DS**, describes information specific to the storage media.
- **Meta information DS**, contains information that usually cannot be extracted from the signal itself.
- **Segment DS**, contains the different segments created by applying different segmentation criteria to the track, alongside additional segmentation information.
- **Auralization DS**, describes different ways for quick listening or "auralization" of contents.
- **Track Linking DS**, describes if the track is linked to another track (i.e. shares the same content descriptions) or any other kind of interaction or relationship between tracks.

Segment Description Scheme

An audio program can be segmented into one or more groups of contiguous samples that share some common property. Each one of such groups is an audio segment. Audio segments can be further segmented into another group of segments, yielding a tree segment. A track can contain different segment trees in order to accommodate different overlapping segmentations according to different criteria (ex. by speaker gender and by background music/not background music).

To each node of the Tree there is attached a Segmentation Criterion DS (except for the Leaves which are not segmented further). The Segmentation Criterion DS is a container for one segmentation criterion, the categories it generates and the segments derived after such a segmentation. The Audio Segment DS is described by a set of DS's similar to the ones used for the Track DS.

Lowest levels of the Segment Tree can be frames or samples.

Region Description Scheme

This description scheme allows the description of the microstructure of a segment of sound, in the form of a sound spectrum, or a series of sub-spectra inside the effective bandwidth of a sound. The region can be linked with a long segment, but also to one as short as a specific frame.

A Region is described by:

- **Audio Spectrum DS**, which describes the main spectral features of the sound.
- **Motion/Deformation DS**, describes the evolution of the Audio Spectrum along a short period of time.

There can be several Region decompositions associated to the same Segment, and it is also possible that several Segments share the same Region Decomposition.

Audio Spectrum Description Scheme

The Audio Spectrum DS allows the specific description of a Region according to its spectral characteristics. The Audio Spectrum DS describes the spectral characteristics of a region and this way, as the Region can be linked to frames or to longer segments the Audio Spectrum DS allows a description of sound ranging from the microstructural level (when the Region DS is linked to a low-level Segment DS as it is the frame), to the macrostructural level (when the Region DS is linked to a high-level Segment DS).

A Spectrum DS consists of a series of specific global descriptors like amplitude, fundamental frequency, spectral range, noisiness, etc. with their statistical

measures, and a *Spectral Shape DS*, that describes the energy profile across the spectrum of the frame.

The Spectral Shape DS describes the shape of the spectrum of an Audio Spectrum. It is described by a Spectral Envelope profile that can be stored in different ways: as a set of LPC or other kind of coefficients, as a list of amplitudes for partials, as Mel-cepstrum coefficients, as Formants, etc. It can also be decomposed into sinusoidal and stochastic components as in spectral models.

3.2 Descriptor Schemes for the semantic level

In the semantic part we describe two broad categories of sonic elements: events and objects. Given the temporal nature of audio events we found that they are mainly linked with segments described in the syntactic part, whereas the audio objects are closely (but not exclusively) linked to the regions described there. Events can be considered as temporal constrained behaviors of sound objects (sound sources). Objects, then, are the subjects of the events.

Event Description Scheme

An event is the temporal behavior of some audio object along or around a certain segment of time. Typical audio events can be: a melody, a musical phrase, a "solo" section, a musical note, different sections of a note regarding its amplitude evolution (attack/steady state/release), an audio fade (in or out), a sentence uttered by somebody in a video, the words that compose the previous sentence, the phonemes that the previous words are made of, non-linguistic utterances made by characters in a video (cryings, shoutings, sighings, etc.).

Audio Object Description Scheme

The main function of an Audio Object DS is describing sound sources. It is possible to distinguish different levels for describing audio objects. The most generic source objects can be: musical instrument, speech voice, environmental sound and sound effects.

4. Conclusions

MPEG-7 is an standardization initiative at the early stages of definition in an area that is not well developed from a technological point of view. Many contributions are needed to make the standard a successful one. We do not claim to have presented a complete proposal in the area of audio descriptors and descriptor schemes, just some ideas that can contribute to the MPEG-7 initiative and other related standardizations.

Part of this work has been done in the context of the European Esprit project 28793 (CUIDAD).

References

CUIDAD. *CUIDAD Working Group Homepage*. http://www.ircam.fr/cuidad.

Foote, J. T.. 1997. A similarity measure for audio classification. In *Proc. AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora*, Stanford, CA, March 1997.

Grey, J. M and J. A. Moorer. 1977. "Perceptual evaluations of synthesized musical instrument tones". *JASA*, 62 (3).

Herrera, P. and J. Bonada. 1998. "Vibrato extraction and parameterization in the Spectral Modeling Synthesis framework". *Proc. DAFX98: First Digital Audio Effects Workshop*. Barcelona, 1998.

Martin, K. D. 1998. "Toward automatic sound source recognition: identifying musical instruments". *Proc. NATO Computational Hearing Advanced Study Institute*, Il Ciocco, Italy, July 1-12, 1998.

MPEG-7. 1999. "MPEG-7: Context, Objectives and Technical Roadmap", ISO/IEC JTC1/SC29/ WG11/N2729V.11. Seoul meeting, March 1999.

Pfeiffer, S., S. Fischer, and W. Effelsberg. 1996. "Automatic audio content analysis". *Proc. ACM Multimedia 96*. Boston, MA. November. pp. 21-30.

Pollard, H. F., and E. V. Janson. 1982. "A tristimulus method for the specification of musical timbre". *Acustica*, 51.

Serra, X. and J. Bonada. 1998. "Sound Transformations on the SMS High Level Attributes". *Proc. DAFX98: First Digital Audio Effects Workshop*. Barcelona, 1998.

Wold, E., T. Blum, D. Keslar, and J. Wheaton. 1996. "Content-based classification, search, and retrieval of audio". *IEEE Multimedia*, pages 27-36, Fall 1996.

Wright, M, A. Chaudary, A. Freed, D. Wessel, X. Rodet, D. Virolle, R. Woehrmann, and X. Serra. 1998. "New applications of the Sound Description Interchange Format". *Proc. ICMC-98*, Ann Arbor, MI. 1998.

Wyse, L. and S. W. Smoliar. 1998. "Toward content-based audio indexing and retrieval and a new speaker discrimination technique". D. F. Rosenthal and H. G. Okuno (Eds.) *Computational Auditory Scene Analysis: Proceedings of the IJCAI-95 Workshop*. Erlbaum, 1998.

Zhang, Tong and C.-C. Jay Kuo. 1998. "Content-based *SPIE's 43rd Annual Meeting - Conference on Advanced Signal Processing Algorithms, Architectures, and Implementations VIII*, SPIE Vol.3461, p432-443, San Diego, July 1998.