Metadata for music and sounds: The Cuidado Project

François Pachet
Sony CSL-Paris, pachet@csl.sony.fr

Abstract:
The IST project CUIDADO is the result of two years of concertation at the European level in the context of the CUIDAD Working Group (Esprit 28793). CUIDADO - led by Ircam, started in January 2000 and will end in December 2003 - aims at developing content-based audio modules and applications using the MPEG-7 media representation standard. The project covers the analysis process (extraction of descriptors), the navigation process (retrieval methods and interfaces implemented in a leading database system with Web interoperability), up to the creative process (consuming and authoring tools) involving content creators and consumers at each stage. The project addresses both the audio (samples) and the music (titles) domains with the assumption that high-level descriptors for music should rely on robust lower level audio descriptors (pitch, energy or spectral features) in order to cover a wide range of applications. This approach matches the needs of record labels and copyright societies for Information management methods for both marketing and protecting their contents. CUIDADO is also a first attempt to go beyond content retrieval by providing an Authoring system using content features for professional musicians and studios.

1. Project rationale and objectives

The Internet explosion has brought a strong pressure to design systems that cope with the complexity of large digital libraries. This is particularly true for the musical world, both in the professional and in the non-professional domain. Millions of music titles are now available in a way or another, leaving potential listeners with the impossible task of managing themselves these huge quantities of music. In the professional domain, recording studios and musicians using digital sound synthesis (that is, most of them) are also flood with terabytes of sounds, and the need for managing them is as urgent and crucial than for music.

The IST funded *Cuidado* project (IST-1999-20194) aims precisely at developing techniques for extracting musical metadata in the large, and validating their relevance in the new context of digital music production and distribution, for building new generation digital sound / music systems. Cuidado is a natural follow-up of the European Working Group *Cuidad* (Esprit 28793), which originally aimed at promoting the development of metadata techniques for music and sound, in the context of the Mpeg-7 standardization effort (Mpeg-7, 1998). Both Cuidad and Cuidado are managed by Ircam.

The project addresses research issues related to feature extraction from the signal, but also integrates the design and implementation of two exemplar systems, one in the music domain (the *Music Browser*), and one in the professional sound production domain (the *Sound Palette*). Cuidado integrates to this end key expertise in various areas:

- Signal processing: Ircam (Paris, France), University Pompeu Fabra (Barcelona, Spain), Sony CSL (Paris, France), University Ben Gourion (Israel),
- Auditory cognition (Ircam),
- Music catalogue management (Sony CSL and Sony Music),

- Database management and representation of large scale collections of data and metadata (Oracle - Spain),
- Development of professional Sound Systems (Creamware, Germany),
- User group validation and dissemination (Artspage, Norway).

There are obvious links between the expected outputs of the Cuidado project and the Mpeg-7 standard. First, results pertaining directly to Mpeg-7 will be proposed to the audio committee of the standard, as was done already with the Cuidad working group (e.g. timbre descriptors, see Cuidad, 2000). Second, the two applications of Cuidado are likely to be among the first Mpeg-7 compatible applications, and will therefore serve as prototypical examples of metadata exploitation in two key areas of digital multimedia.

Cuidado is roughly divided in three main parts: a kernel "server" module containing metadata extractors, and two clients on-line applications. The kernel part concerns strictly the issues of metadata extraction (Section 2). The two applications (Sections 3 and 4) are specifically devoted to the issues of metadata exploitation. Cuidado also contain numerous other significant modules such as music recognition, database management, etc. which are not described here for reasons of space.

2. Metadata Extraction

Metadata extraction is the crucial ingredient in the new chain of digital media production and distribution. It is only if metadata of sufficiently good quality can be extracted automatically that metadata standards and related visions will succeed. Metadata extraction is not a new domain, but has gained attention recently with Mpeg-7 (Philippe (2000), Aigrain (1999)). In the domain of computer vision and image processing, the problems of segmentation, object recognition, scene analysis, etc. have been addressed for a long time, with, arguably, limited success. In the field of audio and music, metadata extraction is, we believe, much more realistic. Music is, from a signal viewpoint, simpler than video, but more importantly, we know more about the nature of the expected descriptors than in the video domain. This is the case, probably, because music has long been the object of sophisticated and successful formalization, and because traditional slicing up of music into various dimensions – melody, harmony, rhythm, expressivity – is commonly accepted and understood by the vast majority of the listening population: there are no equivalent concepts for rhythm, melody or harmony for video…

The extraction and representation of several important dimensions of music have been the subject of reasonably successful approaches. For instance, beat extraction and tempo induction work reasonably well (see, e.g. Scheirer, 1998). Segmentation and music / speech discrimination have been also addressed quite successfully (Rossignol et al. 1998). Pitch extraction was addressed, e.g. by Lepain (1999), although its feasibility on noised polyphonic music remains to be assessed.

One aim of the Cuidado project is to identify other low-level descriptors pertaining to sound and music, and to develop techniques to extract them in a systematic fashion. These low level descriptors include standard features such as spectral centroid, which form the bulk of the Mpeg-7 audio standard, as well as, for instance, *segmentation* (cutting a piece of music into different consistent fragments), timbre descriptions (with an emphasis on non stationary sounds such as percussions), voice categorisation, etc.

Higher-level descriptors will also be addressed, in particular for applications in Electronic Music Distribution (*EMD*). The objective here is to characterizing music titles in their entirety for cataloguing and search / retrieval. Targeted descriptors include rhythm structure (with preliminary results at Sony (Gouyon et al., 2000), global timbre characterization, energy descriptors, and in general all the descriptors which make sense on a global title level and may be useful for EMD applications.

It is important to note here that there is indeed a scientific assumption underlying this goal: the assumption that current techniques in signal processing (mainly spectral analysis, information theory and statistical analysis) are sufficient ingredients to perform these extractions. At least, the Cuidado project will experiment with these techniques in the time span of 3 years, with a high chance of success for most of the descriptors envisaged.

3. Music Browser

The Music browser is a system able to provide various content-based access methods for large catalogues of popular music. This system embodies the different visions of the Cuidado consortium – and the Sony team in particular, concerning content-based access (Pachet, 2001). First, the system will manage a set of high-level descriptors, as outlined in the preceding section. These descriptors will be designed to yield robust, grounded similarity measures between music titles. Each of these descriptors should in principle yield basic “query by Xing” services: query by humming, if polyphonic pitch extraction is successful enough, but also query by drumming, and all incarnations of “query by examples” along the corresponding musical dimensions.

Other music descriptors will be investigated, which cannot be, on first approximation, extracted from the signal. These descriptors include typically genre and sub genre (Pachet, 2000), and other social-related information such as provided by collaborative filtering techniques (Cohen et al, 2000). We investigate data mining co-occurrence techniques (Pachet and Westermann, 2001) to extract these descriptors, on large corpora of music data (radio programs, web sites, sampler catalogues).

Descriptors are useless without retrieval methods able to exploit them: we develop music access methods that exploit descriptors to provide high-level means of browsing through music catalogues. These methods consist, in particular, in proposing user to search for fully-fledged music programs (seen as temporal sequences of titles), rather than for actual titles, and were already experimented successfully in small catalogues (Pachet and al., 2000). The techniques used range from complete combinatorial search using constraint satisfaction techniques to more recent incomplete, but faster methods using local search.

Finally, the music browser will integrate user interfaces designed to handle most of the known music listening behaviours, from the genre-oriented, focused types to open and exploratory behaviours. User validation is conducted from the very beginning of the project on existing prototypes to ensure that the whole range of techniques developed is actually relevant, and allows to efficiently match arbitrary music tastes to music.

4. Sound Palette

The Sound Palette system draws from Ircam’s *Studio On Line* project (Ballet, 1999), and consists in developing metadata extraction techniques for describing and managing large collections of sounds, in the context of music production environments. Although there is no precise definition of *sound* (as opposed to music titles), in Cuidado sounds are, in principle, audio items to be eventually integrated in a music composition. As for music titles, descriptors for sounds fall into two categories: low-level, well-defined, and probably not very useful descriptors (with the prototypical spectral centroid), and higher-level, much needed descriptors for categorizing sounds in general, such as addressed for instance by Wold at al., 1996, but tailored for music production applications.

The Sound Palette wil serve to purposes: 1) content-based retrieval (search by perceptual similarity, automatic classification techniques, use of textual descriptors), addressed by Ircam Analysis/synthesis, Ben Gurion University, and 2) high level sound editing based on descriptors (Xavier Serra' s team).

The resulting system should allow musicians to find sounds quickly, using high-level search tools, including sequence-based as for the music browser, and also possibly allow groups of musicians to share efficiently sounds and sound repositories.

5. Conclusion

Cuidado is a very exciting project for a number of reasons. It is the first time that a substantial collective effort will be made to close the loop from music/sounds to users, using the whole palette of signal-processing and computer science technologies available. As a side effect, it is certain that a number of basic research issues will be reconsidered successfully thanks to the nature of the applications, and possibly new basic research issues will pop up (for instance it is expected that issues related to non stationary sounds will receive much attention). Second, the applications envisaged are, in fact, long overdue, both for the Music browser and the Sound palette: the consortium will try to capture the energy of these willing users to ensure a strong adequacy between the desires of users and the outcomes of the project.

References

Aigrain, Philippe (2000) New Applications of Content Processing of Music, Journal of New Music Research, 28:4.

Ballet Guillaume, Borghesi Riccardo, Hoffmann Peter, Lévy Fabien, "Studio Online 3.0: An Internet "Killer Application for Remote Access to IRCAM Sounds and Processing tools", Actes du colloque Journées d' Informatique Musicale, 1999

Cohen, W., Fan, W. (2000) Web-Collaborative Filtering: Recommending Music by Crawling The Web, 9th International World Wide Web Conference, Amsterdam.

Cuidad Esprit working Group: http://www.ircam.fr/cuidad. In particular, see W4004, Text of ISO/IEC FCD 15938-4 Information Technology - Multimedia Content Description Interface - Part 4 Audio, Singapore March 9, 2001.

Cuidado IST project main page: http://www.ircam.fr/cuidado

Doval B. (1995) Fundamental frequency estimation of musical sounds using statistical learning. In *International Symposium of Musical Acoustics, Dourdan*.

Gouyon, F. Delerue, O. Pachet, F. (2000) “Classifying percussive sounds: a matter of zero-crossing rate ?” Digital Audio Effects Conference, Verona (It).

Lepain, Philippe (1999) Polyphonic Pitch Extraction from Musical Signals, Journal of New Music Research, 28:4.

Mpeg-7 standardization process, http://www.darmstadt.gmd.de/mobile/MPEG7/

Pachet, F. Roy, P. Cazaly. D. (2000) “A Combinatorial approach to content-based music selection”. IEEE Multimedia, pp. 44-51, March 2000.

Pachet, F. Cazaly, D. (2000) “A Classification of Musical Genre”, Content-Based Multimedia Information Access (RIAO) Conference, Paris.

Pachet, F. Westerman, G. (2001a) Music Similarity for EMD, submitted to WedelMusic 2001, Firenze (It).

Philippe, Pierrick (2000) Low-level musical descriptors for MPEG-7, *Signal Processing: Image Communication*, 16 181-191.

Rossignol, Stéphane & al (1998) "Features extraction and temporal segmentation of acoustic signals", Proc. ICMC.

Scheirer, Eric D. (1998) “Tempo and beat analysis of acoustic signals”, JASA, 103(1).

Wold, Erling; Keislar, Thom; Wheaton, James (1996) Content-Based Classification, Search, and Retrieval of Audio, IEEE Multimedia, 3:3, pp. 27-36.