The CUIDADO Project : New Applications based on Audio and Music Content Description

Hugues Vinet*, Perfecto Herrera‡, François Pachet◊
***IRCAM, ‡UPF, ◊SONY CSL**

Objectives and methodology

The CUIDADO project (*Content-based Unified Interfaces and Descriptors for Audio/music Databases available Online*) aims at delivering new applications based on the systematic use of audio and music descriptors, which address new needs of the professional and consumer audio/music chain. This approach falls within the scope of the MPEG-7 standardization process, in which CUIDADO participants have been already actively involved [Peeters00]. In this framework, descriptors and descriptor schemes combine knowledge bases and numerical features of the audio contents, and their use is intended for several classes of applications. First, as metadata in audio/music databases, enabling content-based retrieval functions, which would not be feasible, due to data volumes, by directly accessing audio samples. Secondly, for high-level manipulation of audio/music contents, considering that operations on the sound material must be specified in a way as close as possible to the user's mind in his relation to music, and not oblige him to assimilate low-level, meaningless implementation-driven data structures. More generally speaking, the design of appropriate descriptors makes possible new application models for accessing and manipulating audio contents and represents a major opportunity for the development of the music production and distribution industries. However, in order to come up to effective results, several methodological considerations, which form the basic assumptions on which the project is built, must be taken into account:

i) the necessity of a top-down approach, aiming at deriving *high-level descriptors*, which are to be the basic knowledge structures accessible by the user in target applications. These knowledge structures must be designed in a way such that they are adapted to targeted application functions and consistent with the user's own cognitive structures.

ii) the interest of combining this top-down approach with a bottom-up approach, which extracts *low-level descriptors* from an automatic analysis of the audio signal. The objective is to automatically compute required high-level descriptors, avoiding manual input when possible. The main related issues, discussed in [Herrera02] in the context of sound classification, are how to choose a set of relevant low-level descriptors and how to map targeted high-level descriptors onto them, by using appropriate machine learning techniques.

iii) designing and validating descriptors in laboratory conditions may however not suffice for real-world applications. In order to enable a complete assessment procedure, it is necessary to build fully-functional applications prototypes, which will take into account all technical constraints and man-machine interaction issues. Machine learning functions enable the system to learn the user's knowledge structures; symmetrically, the design of the user interface must be considered as a key factor for the user to be able to assimilate the system knowledge structures.

iv) such an ambitious program requires the combination of a multi-disciplinary scientific and technological expertise, which has been gathered for the CUIDADO project in an international consortium including approximately 40 researchers and engineers from IRCAM (coordinator), IA-UPF, SONY-CSL, Ben Gurion University, and industrial partners (Oracle, Creamware, Artspages). The project is supported by a grant from the European Commission (Information Society Technologies Program).

In order to fulfill these objectives and methodology, the project is organized around the development of two target applications, the Sound Palette, and the Music Browser, which address different needs. These applications are presented in the next sections.

The Sound Palette

The Sound Palette is an application dedicated to professional users in music and audio production. It offers audio sample management and editing features based on sound content description.

Sample management features

The notion of audio sample is taken here in its usual denomination, i.e. sounds of short duration (typically less than 20 seconds), which contain individual events, such as single instrument notes. Actually, the system can handle samples with more complex contents, but the main assumption made on related descriptors is that they will be computed globally for the whole sound, without any segmentation/separation pre-process.

The first available application using content-based descriptions for sample management has been SoundFisher by MuscleFish [Keislar99]. The Studio Online project realized at IRCAM [Wörmann99], provided an online server giving access to a 117,000 instrument sample database, where high-level retrieval functions were experienced: browsing through a specifically designed sample hierarchical

taxonomy (including contemporary playing modes) and a query by example interface, relying on an elementary perceptual similarity distance resulting from studies on timbre spaces. A more general descriptor scheme for instrument timbre was proposed and integrated in the MPEG-7 standard [MPEG-7]. Targeted features for the Sound Palette include various enhancements, in particular by enabling the management of the user's own sounds, and not only access to a fixed database as in Studio Online. They combine the following functions:

- classification tools, which help the user to organize his sounds in classes, and provide automatic classification solutions when introducing new sounds. Classes can be learned from arbitrary user categories (as long as they refer to the sound content), and can also be trained from more objective classification schemes, such as sound sources, or morphological characteristics (such as pitch or amplitude envelopes, grain, etc.).
- management of a global sample repository, shared among several users, containing both a reference set and user samples, and available as a set of online Intranet/Internet services.
- sample retrieval functions, combining criteria used for classification (limitation of the search set) and query by example functions, based on similarity distances, computed between each pair of samples across class boundaries.

The first tests of automatic classification algorithms, based on static low-level descriptors (computed for the whole sound duration), performed on standard instrument classes, show promising results [Peeters02]. New, more robust, perceptual similarity functions are also proposed, based on recent results of a meta-analysis of main existing studies on timbre spaces [McAdams]. Further developments foresee the use of dynamic descriptors, which model the evolution of various sound features over time, and the extension of proposed search functions through the use of textual attributes, which qualify various aspects of the sound contents.

Two versions of the Sound Palette are developed in parallel :

- an "online version", which includes all above functions, and is experimented as a common sample database for all musical productions at IRCAM.
- an "offline version", integrated in an audio production workstation such as the Creamware Scope system, which is restricted to the management of the user's samples on his local PC, but also includes sample editing features described in the next section.

Sound editing features

The Sound Palette can handle three different kinds of sound materials : sound samples, music monotimbral phrases, and rhythm tracks (as for example the "drum loops" that are ubiquitous in current popular music).

Deriving a multi-level music description of these materials and populating large databases with them paves the way for new navigation, editing and transformation possibilities. The functionalities are even more enhanced when a synthesis engine is allowed to play some role in the general architecture of the system. Provided this system, we can talk about "Content-based sound edition", a concept that has much in common with old proposals of "intelligent sound editors" [Chafe89].

Navigation in the Sound Palette editing environment may use note, phrase or instrument labels instead of the handwritten markers that are usually mandatory in standard editors. "Skip to next note" or "Skip to next snare hit" provide better location than spotting the cursor somewhere "there" by visually inspecting the waveform and scrub-listening to it.

Automatic segmentation and labeling of content also allows multiple selections of those segments. Once they are selected, it is possible to perform editorial actions on all of them simultaneously. For example, after automatically describing an electronic bass phrase with note and sub-note descriptors, it is possible to select a portion of their attacks and chop them (a well-known technique in certain popular styles). This is done in a compact way instead of manual-visual selection and change of each one of the portions to be edited.

Creation of Midi maps from an audio rhythm loop is another convenient function that cannot be completely achieved with current commercial software devoted to this type of sonic materials (e.g. Recycle!). A Sound Palette rhythm description contains not only information about onset location of sounds and basic pulses that underlie in the track but also generic labels for the percussion instruments. This way, a Midi "recreation" of an example drum loop can be set up for using our owned drum kit samples, instead of some illegal CD-lifted fragment. Thanks to a perceptual similarity-based search function, drum samples used in the recreation can even be perceptually very close to the original ones (provided a large sound database, indeed!), not only the same in terms of category.

As editing and transforming sounds become operations with no clear boundaries between them, transformation comes into play even in editorial tasks such as concatenating two rhythmic patterns for using the composite as a song building block. Let's suppose that pattern A has a global tempo of 100 BPM and pattern B scores at 80 BPM, and we want the composite at 93 BPM. Rhythm tempo matching can be achieved by conveniently time-stretching the original patterns. Seamless and continuous fusion between different-tempo rhythm

patterns is also possible without bothering the user with tempo or stretch parameter adjustment.
Variation generation is a must in the composer armamentarium, and the Sound Palette implements some options for that. Traditional variations on a musical phrase by tonality change of a phrase are achieved through transposition of specific notes from the original phrase. Rhythm tracks can also be altered by swapping a specified instrument by another, or by muting the specified one.

The Music Browser

The Music Browser intends to be the first content-based music management tool for large music catalogues. This application targets all the actors of the music distribution industry: labels (systematic exploitation of catalogues), distributors (personalized music playlists), broadcasters (music programs that match user's tastes). Several music browsers have been made public recently. The ones that propose some sort of content-based approach (MoodLogic, Relatable) suffer from two main drawbacks:

- The metadata used is entirely manual, thus very expensive to build and maintain.
- There is no facility for automatically building music playlists.

The CUIDADO demonstrator implements the whole processing chain, from the music data (signal, Midi, or reference information) to the user, with as much automatic processing as possible. The aim is to build concrete evidence that content-based music management does allow efficient personalized music access and brings substantial added value in comparison to existing music distribution schemes.
Technically, the main challenge of the Music Browser is to be useful for the largest possible set of music behaviors. Preliminary experiments show indeed that user behaviors exposed to large music catalogue vary greatly. Some users want to browse using top-down genre categories, others are example-oriented. Some know exactly what they want, others want only titles they do not know (see [Pachet CACM, 2002]). The Music Browser implements many browsing mechanisms, including:

- *Descriptor-based* search: "I want titles with fast tempo, high energy, in the Rock generic genre",
- *Similarity-based* search: "I want titles which are close to a given set of titles". The similarity may be picked up from a list of several similarity relations, each one bearing different semantics, e.g. cultural similarities, metaphoric similarities, artist-based similarities, etc.
- *Global search*: "I want a playlist with instrumental music at the beginning, dance music at the end, with increasing tempo values, and more than 60% of female singing voice".

Unary descriptor extractor module

This module implements top-down signal processing techniques that extract global, unary information about music titles, considered here as 3 to 4 minute signals. Targeted descriptors are of 3 sorts:

1) *Rhythmic*. This category includes tempo/beat, as well as rhythmic information. In particular, we have focused on percussive rhythm, i.e. rhythm produced by repeated occurrences of percussive sounds [Gouyon00].
2) *Energy-based*. These descriptors aim at extracting rough categories of perceptual energy. A typical example is the intuitive perceptive difference between a hard rock piece and an acoustic guitar folk song.
3) *Timbre-based*. These descriptors aim at extracting a global representation of the overall timbre of the song. The resulting descriptor serves as a basis for establishing timbre similarities between songs.

Similarity analysis modules

These rather low-level descriptors described above produce accordingly low-level similarities. There are other means of extracting higher-level similarity relations for music. A well-known technique for producing similarities is collaborative filtering (Firefly). This technique builds similarities between titles based on similarities between user profiles. We have investigated more precise techniques based on data mining, applied to corpuses of textual information. The corpuses include album playlist collections (e.g. CDDB), radio program listings, and general search engines (e.g. Google). The techniques we use are based on co-occurrence. We have shown in [Pachet01] that the extracted similarity is meaningful, non-trivial, and complementary to the similarities extracted through low-level descriptors.
These similarity relations are exploited in the music browser for similarity-based search, as well as for proposing music discovery schemes. Through a simple user interface, the user can quickly explore regions of the catalogue which are either very close to his set of preferred titles, or very far (see [Pachet02]) for a prototype implementing varying-length music catalogue exploration).

Playlist generation module

As mentioned in the preceding section, music titles are rarely accessed and listened to in isolation. We have proposed in [Pachet et al., 1999] to consider sequence generation as the main paradigm for music retrieval, instead of the standard title-based approach. One of the main interests of sequence

generation is that it allows users to explore efficiently the catalogue without having to learn the underlying descriptor ontologies. The properties of music playlists include continuity (a playlist which is continuous, tempo-wise, or style-wise), cardinality (at least 60% Funk titles), and distribution (maximum 2 titles by the same artist).
For instance, a 10 title playlist can be specified with the following properties:

- All titles are different,
- Increasing tempo,
- Three cardinality constraints on genre:
  - less than 50% World,
  - exactly 4 Rock,
  - more than 50% Folk.

Finding solutions for large catalogues is inherently NP-hard. We have developed an incomplete but efficient algorithm that scales up to over 200,000 titles [Aucouturier02], and handles arbitrary complex properties.

Web Music Monitoring System

This module, integrated in the Music Browser, is specifically intended for copyright societies. It performs the identification of unknown music excerpts, e.g. found on the Web, with a reference copyrighted title database. The identification principle is based on signal comparison. It relies on audio signatures, stored as metadata, which capture relevant information by encoding original signals with high compression rates.

Architecture

A common hardware and software architecture has been designed and set up for both applications, and provides all required services, including : massive data storage, audio streaming, file uploading and downloading, audio file, metadata and user data management, middleware, etc. All used protocols are chosen to be as standard as possible; access to online services is provided from a variety of client terminals, through the use of HTML/XML when possible, otherwise Java protocols for client interfaces. The system, based on a 3-tier architecture, is shown in Figure 1.

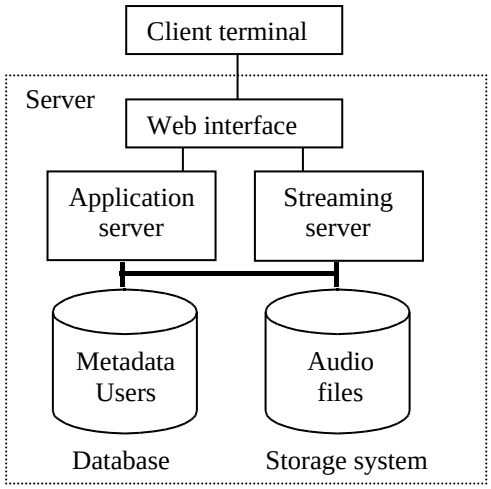

Figure 1 – System architecture

References

[Aucouturier02] Aucouturier, J.-J. and F. Pachet Scaling up Music Playlist Generation, submitted to *IEEE International Conference on Multimedia Expo*, 2002.

[Chafe89], Chafe C., B. Mont-Reynaud and L. Rush (1989) "Toward an Intelligent Editor of Digital Audio: Recognition of Musical Constructs" In C. Roads (ed.), *The Music Machine*. MIT Press, Cambridge, MA.]

[Gouyon00] Gouyon F, O. Delerue, F. Pachet, "Classifying percussive sounds: a matter of zero-crossing rate?" *Digital Audio Effects Conference*, Verona (It), December 2000.

[Herrera02] Herrera P., G. Peeters, S. Dubnov, "Automatic Classification of Musical Instrument Sounds", *Journal of New Music Research*, 2002 (*to appear)*

[Keislar99] Keislar, D., T. Blum, T., J. Wheaton, & E. Wold,). "A content-ware sound browser". *Proc. of the International Computer Music Conference*, ICMA, 1999.

[McAdams] McAdams S, S. Winsberg, "A meta-analysis of timbre space. I : Multidimensional scaling of group data with common dimensions, specificities, and latent subject classes" *Journal of the Acoustical Society of America*, (*in preparation)*

[MPEG-7] ISO/IEC FCD 15938-4 Information Technology - Multimedia Content Description Interface - Part 4 Audio.

[Pachet99] Pachet F., P. Roy, D. Cazaly, "A Combinatorial approach to content-based music selection", *Proc. of IEEE International Conference on Multimedia Computing and Systems*, Firenze (It), Vol. 1 pp. 457-462, 1999.

[Pachet01] Pachet, F., G. Westerman, D. Laigre, (2001) Musical Data Mining for Electronic Music Distribution, *WedelMusic Conference*, Firenze (It), Nov. 2001.

[Pachet02] Pachet F., "Electronic Music Distribution: The Real Issues", *Communications of the ACM*, to appear, 2002.

[Peeters00] Peeters G., S. McAdams, P. Herrera, "Instrument sound description in the context of MPEG–7", *Proc. of the International Computer Music Conference*, ICMA, 2000

[Peeters02] Peeters G., P. Tisserand, "Selecting signal features for instrument sound classification", *Proc. of the International Computer Music Conference* ICMA, 2002 (*Submitted*)

[Wöhrmann99] Wöhrmann, R., G. Ballet, "Design and architecture of distributed sound processing and database systems for web-based computer music applications", *Computer Music Journal*, vol 23, Number 3, p.77-84, 1999.