Room Models and Object-Oriented Audio Coding: Advantages and Applications

Giorgio Zoia
Signal Processing Institute, EPFL, Switzerland; Email: Giorgio.Zoia@epfl.ch

Introduction

Coding and transmission of high quality, flexible and interactive acoustic scenes are gaining a more and more relevant place not only in audio, but especially in multimedia and virtual reality modeling. A realistic and immersive sound experience is recognized as fundamental for a satisfactory perception of a virtual space. At the same time, a consistent representation of a three-dimensional audio environment still represents a challenge. In fact, the amount of information to be stored and/or transmitted, the useful size of the listening area, the number of enabled degrees of interaction, the overall computational complexity and quality of sound are often conflicting requirements.

To give some examples, almost all the existing methods for a 3D volume rendering solution, i.e. one that generates an accurate and realistic representation of the original sound field in the entire listening space, require a relatively high number of channels and/or loudspeakers, while cheaper solutions in terms of overall information often lack in quality, or alternatively they rely on very limited (in terms of volume of the listening space) rendering configurations or devices, such as headphones in the worst case. In another typical situation, it is true that more or less artistic forms of sound post-production can offer extremely satisfactory results in terms of quality and a decrease in the number of the several channels that are recorded and stored in a studio session; but at the same time any sound post-processing technique can be seen as a dramatic shrink in the degree of potential interaction for the end user. If in some situations this is even desirable, in other cases it constitutes a limitation to the flexibility and to the usability of the sound content as a communication mean.

In the remaining of this paper it is shown how a hybrid sampled (or natural) and structured (or model-based) sound representation can provide a very effective coding solution, where all the above requirements can be satisfied at best through a compact and flexible technology. Moreover, this systematic approach to audio coding brings as a potential advantage a realistic solution to the problem of future coding and transmission formats for 3D interactive acoustic scenarios, for which a clear distinction between *coding* and *rendering* is desirable and necessary.

Towards a Hybrid Audio Coding Scheme

Mono, stereo and 5.1 sound formats, the ones that can be normally "purchased" in stores on CDs and DVDs, present a very strong relationship with the rendering configuration. When a sound engineer mixes single tracks to a master track, this is often done for one particular configuration of the speakers, often according to the target application (music CD, movie soundtrack, radio mix and so on). That is, the way of coding audio is mostly channel-oriented.

Object-oriented Coding

In an object-oriented coding approach, the several sources (single- or multi-channel) are represented separately instead of being mixed into a suitable number of channels; if these sources can be further associated to appropriate models of the room acoustics (or, more in general, models of the acoustic environment), some important benefits can be obtained. First, the representation of the audio content is independent from the loudspeaker configuration; moreover, it is possible to reduce the overall amount of information in comparison to multi-channel coding formats, in principle without loss of quality; last, the listener may be able to interact with the content in a more natural and flexible way.

Back to the simple coding schemes considered at the beginning of this section, if e.g. a clarinet is coded as a single source, and not as a stereo pair of channels or as a 5.1 set of channels, and in addition its location in the space, its directivity, its intensity and so on are coded with the (mono) track, the representation of the clarinet is now independent from the number of channels to be used for rendering: a suitable decoder, belonging to the domain of the rendering equipment, will take care of translating the object-oriented coding into a channel-oriented rendering. The same can be done, in principle, for any other instrument and voice that may be part of the sound work.. This approach is represented in Fig. 1 for a 3-objects to 4-channels case.

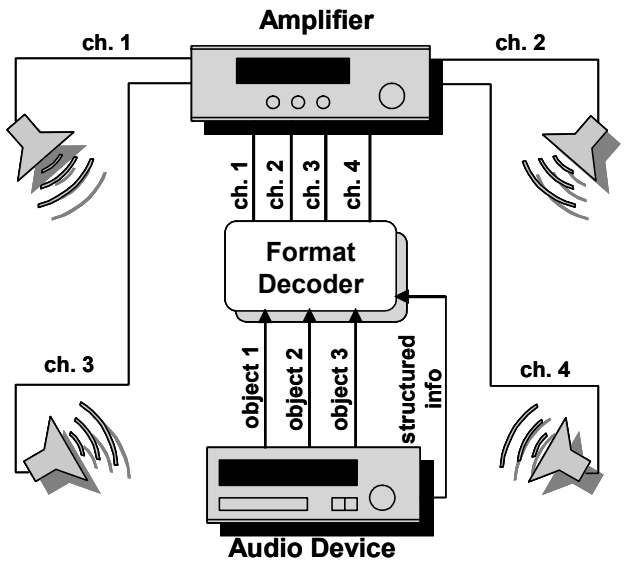

Fig. 1: An object-oriented coding format maps one object to a virtual position, independently from the rendering configuration

A problem may arise when there is more than a reasonable number of objects. In fact, in an object-oriented hybrid structured and natural sound coding, a clarinet or a drum are objects, but a "virtual speaker" (the actual object of which the musician or the sound engineer think) is an object as well. A stereo mix, for instance, is composed of two one-channel objects with their location, directivity diagram, intensity and so on. At the rendering side, the decoder will take care of placing all the objects in the best place it can, according to the capability of the system, the number of speakers, the eventual room compensation requirements, etc. In this way it is not difficult to imagine the most important elements of the *scene* coded independently in full quality and interaction potential, whereas the remaining instruments and voices will constitute the

sonic background that can be produced by a conventional mixing technique.

The MPEG-4 Hybrid Coding Scheme and Room Acoustic Models

The MPEG-4 Audio and Systems standards jointly allow, for the first time in a normative context, an object-oriented source coding with associated physical room descriptions and perceptual features.

In MPEG-4 there are two fundamental layers that are used to code the multimedia content. At the first layer, which can be called here for clarity the *media* layer, audio-video objects are delivered to a terminal as elementary streams and they are converted back to samples (or pixels) by several instances of the corresponding decoders. The sources are then made accessible at the upper layer, the *systems* layer (or more intuitively the *scene description* layer), composed by a structured scene description and by dynamic update mechanisms, which are in charge of the manipulation of the parameter space of the model.

In more details, MPEG-4 provides several audio *Profiles* that combine conventional but high-quality schemes like perceptual coding (e.g. for dry tracks or, in general, traditional recordings) with parameterized coding extensions, structured voice coding (e.g. CELP codecs) and sound synthesis within a highly structured post-processing subtree.

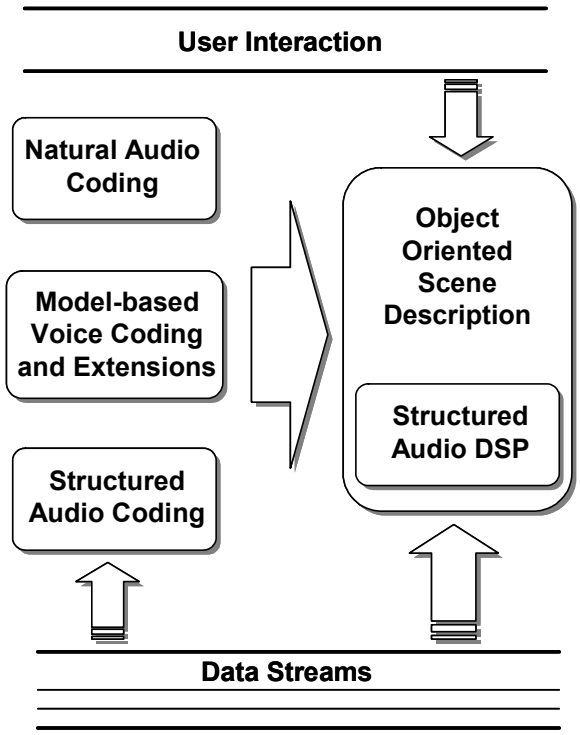

Fig. 2: Structure of the MPEG-4 Coding Scheme

At the systems layer, BIFS (for BInary Format for Scenes) is a mark-up language for scene descriptions; it is a language very similar in its hierarchical structure to VRML, but with an innovative audio-specific processing subtree, often referred to as AudioBIFS; functionality of AudioBIFS ranges from simple mixer or delay *nodes* up to advanced spatialization schemes based on geometrical and perceptual information, so that a complete virtual audio environment can be modeled quite precisely (for exhaustive details see [1], Chapter 12). By this approach an object-oriented coding scheme like the one described in the previous section can be easily implemented to support any combination of real objects and channels. The MPEG-4 coding scheme is summarized in Fig. 2.

MPEG-4 extends simple virtual-reality models of previous acoustic scene descriptions by the inclusion of two advanced and robust techniques to create virtual audio environments. The first technique is *physical*: modeling of the acoustic environment is based on the physical reality defined by the visual scene. The second is *perceptual*: creation and modification of environmental sound characteristic is based on perceptually significant parameterization.

A physical model of an acoustic environments is a sound processing modeled with the purpose that the acoustic effect corresponds to the visual scene. This technique requires models for individual sound reflections of the walls, models for sound propagation through objects, the simulation of air absorption, and the rendering of late diffuse reverberation, in addition to the 3D positional rendering of source locations. This type of environmental spatialization is sometimes referred to as auralization or *virtual acoustics* [2].

Audio spatialization can also be approached from a non-physical point of view, investigating the perceptually relevant quality of spatial audio and room acoustics. This approach is intended mainly for applications where the environmental response does not have to correspond to the visual environment, but where high-quality virtual room acoustic effects are nevertheless desired. The perceptual parameter space standardized in MPEG-4 is derived from the IRCAM *spatialisateur* [3].

CARROUSO

The hybrid, object-oriened coding scheme discussed above is exploited, other than in more general research activities, as an MPEG-4-based supporting technology for the European IST project CARROUSO, which relies on wave field synthesis (WFS) technology [4] for what concerns rendering. CARROUSO will provide a new technology that enables to transfer a sound field, generated at a certain real or virtual space, to another usually remote located space, with full interactive control of the same sound field. MPEG-4 is probably the only international standard to date being able to provide the flexible and powerful coding scheme that is necessary to efficiently support WFS requirements and sound quality, where a huge number (normally 48, at least) of loudspeakers is necessary to properly render a high-quality holophonic acoustic environment (i.e. no sweet spot) and without spatial aliasing [5].

---

[1] F. Pereira and T. Ebrahimi (editors): The MPEG-4 Book, IMSC Press, Prentice Hall, 2002.

[2] G. Gardner: Reverberation algorithms, in Applications of Digital Signal Processing to Audio and Acoustics, M. Kahrs and K. Brandenburg, Eds. New York: Kluwer Academic, 1998, pp. 85-132

[3] Jean-Marc Jot, Olivier Warusfel: Spat: A Spatial Processor for Musicians and Sound Engineers, in Proceedings of the International Conference on Acoustics and Musical Research, CIARM '95, Ferrara, Italy, May 1995

[4] A.J.Berkhout, D. de Vries, P. Vogel: Acoustic Control by Wave Field Synthesis, Journal of the Acoustic Society, vol. 93, pages 2764-2778.

[5] U. Horbach, M. Boone: Future Transmission And Rendering Formats For Multichannel Sound, in Proceedings of the 16th AES Conference on Spatial Sound Reproduction, Ruovaniemi, Finland, April 1999.