Virtual Acoustics Rendering in MPEG-4 Multimedia Standard

Riitta Vääränen and Jyrki Huopaniemi
Helsinki University of Technology
and Audio Signal Processing
and Nokia Research Center
Helsinki, Finland
Espoo, Finland
riitta.vaaraninen@hut.fi
jyrki.huopaniemi@nokia.com

Abstract

This paper reviews the spatial audio capabilities in MPEG-4 scene description. The advanced sound features such as modeling of the acoustic environment in interactive virtual reality applications introduced in the second version of the standard are dealt with. Parametrization of the acoustic environment as well as efficient DSP implementation of immersive 3-D audio environments are discussed.

1 Introduction

MPEG-4 is a standard for interactive multimedia applications and it specifies coding and presentation of audiovisual objects [1]. Like the former MPEG standards, MPEG-4 audio specifies various coding methods for natural audio, but in addition it brings about a scheme where synthetic sound can be coded and transmitted in a structured and parametric form [2]. Furthermore, the presentation of natural and synthetic sound in MPEG-4 can be controlled using novel techniques included in the scene description language of the standard [3].

In MPEG-4, the various audiovisual streams are dealt with in an object-oriented and hierarchical way. Thus the content which is perceived at the terminal may be composed of various audio and visual streams which can be time-dependent, synchronized, and placed in a defined position in a 3-D virtual space, and that may also be interactively affected by the user. The Systems part of the standard specifies a set of synthetic audio and visual objects that may be positioned in a 3-D space [4]. These objects are described in a scene description language called Binary Format for Scenes (BIFS). BIFS encompasses the current Virtual Reality Modeling Language (VRML2.0) specification, which can be used to hierarchically build up interactive virtual 3-D scenes, and where also simple 3-D sound source models are provided [5]. Furthermore, MPEG-4 version 1 expands the VRML audio capabilities with new mixing, manipulation and processing capabilities for natural and synthetic audio [3].

In the upcoming second version of the MPEG-4, the 3-D sound model is extended to allow for more enhanced modeling of sound sources. Also more detailed modeling of the sound propagation in the medium between the source and the listener is introduced, taking into account the propagation delay between the source and the listener, and the objects affecting the sound path [6].

This article deals with the 3-D sound modeling capabilities present in the Scene Description language of MPEG-4. First an overview is given on sound environment modeling, and it is explained how in the forthcoming second version of MPEG-4 the various components of the sound environment are represented. Then, the hierarchical and object-oriented composition of audiovisual objects in MPEG-4 is presented. After that, a DSP implementation for the above system is described, and issues dealing with efficient and time-variant audio processing are discussed. Finally, application targets of the presented system are considered.

2 Parametric Presentation of 3-D Sound Environment

Intelligent parametrization of acoustic environments is an important task for efficient, real-time virtual reality rendering. Essentially, the goal in a virtual acoustic simulation is to produce a binaural room impulse response (BRIR), which corresponds to the sound heard at the ears of a listener in a defined space. Virtual acoustic processing is normally divided into three parts: Source modeling, Room modeling, and Listener modeling.

Source modeling consists of methods, which produce and add physical character to sound in an audiovisual scene. Both natural and synthetic audio may be considered as inputs to the scene description. In MPEG-4 this can be carried out using the audio coding toolset [2]. When giving physical properties to sound sources, their directional radiation characteristics should also be taken into account and parametrized. Typically sound sources radiate more energy to the frontal hemisphere whereas sound radiation is attenuated and low-pass filtered when the angular distance from the on-axis direction increases. Methods for parametrization of sound source directivity have been presented in, e.g., [7].

Room modeling is concerned with investigating sound propagation behavior in acoustical spaces. The ray tracing and the image-source methods are the most often used modeling techniques, when
sound reflections are taken into account. When a response of a reverberating space is auralized in real-time the limited calculation capacity often calls for simplifications, modeling only the direct sound and early reflections individually, and the late reverberation by recursive digital filter structures. The parameters related to room acoustics that should be transferred in a virtual acoustic system are can be divided into global and local ones. The global parameters include the room volume, absorption area, reverberation time. The locally active parameters are the properties of transmitting medium and the boundary surfaces and objects in the room. These parameters include the speed of sound, air absorption, material reflection and transmission transfer functions, and the diffusion and diffraction of sound from the surfaces.

In listener modeling, the properties of human spatial hearing are considered. Simple means for giving a directional sensation of sound are the interaural time and level differences (ILD and ITD), but more accuracy and immersiveness can be gained with head-related transfer function (HRTF) based processing. The HRTF that models the reflections and filtering by the head, shoulders and pinnae of the listener captures the full three-dimensional static sound localization cues. Listener-related parameters are the direction of arrival of the direct sound and the reflections, and the properties of the listener (desired accuracy of HRTF modeling, listener database) to be used for spatial sound reproduction. The listener properties are closely linked to the reproduction method. If binaural reproduction is used, HRTFs are normally required for high-quality synthesis, but for discrete multichannel reproduction the binaural filtering is not needed. It is clear that a general-purpose virtual reality language should not specify any particular method or set of filters to be used for sound reproduction. Therefore, the parametrization of output format and binaural filtering is not incorporated in VRML2.0, MPEG-4 and related scene description languages.

The previously presented parameters describe the physical behavior of sound in rooms. From the modeling point of view, this physically-based approach is the most natural one, since an accurate reproduction of the physical room acoustic mechanisms will yield a natural audible result [7]. However, another viewpoint to the problem is to look at the perception of sound in rooms. This perceptual approach aims at finding subjective criteria (such as presence, warmth, brilliance, envelopment, heaviness, liveness) for describing the room acoustical quality, and linking these parameters to the actual rendering process [8]. Both of these modeling methods have been proposed for MPEG-4 version 2, but in the following sections we only concentrate on the physical modeling approach.

3 Audiovisual scene description in MPEG-4

MPEG-4 BIFS is used to build up audiovisual virtual worlds of hierarchically structured objects. A BIFS scene graph has a tree-like structure which consists of nodes of which the leaf nodes are apparent (visible or audible) to the user, and the nodes above them are used for, e.g., grouping and positioning them in the audiovisual scene. In the following, audio composition and effects processing capabilities of sound in MPEG-4 are discussed, and the 3-D presentation of sound in existing specifications of audiovisual scenes is shortly reviewed.

In BIFS there is a part which specifies how the decoded audio streams are presented to the user at the MPEG-4 terminal. This part of the scene description is called AudioBIFS, and it is used for audio composition by combining the different audio streams to a single audio track in a desired way. AudioBIFS nodes provide functionalities such as mixing, switching between different streams or channels, delaying the sound for obtaining audiovisual synchronization, and buffering samples of sound for its reuse [3]. Also, it is possible to take advantage of the effects processing functionalities provided by the Structured Audio part of the MPEG-4 audio standard [2].

Since MPEG-4 specifies the same set of nodes which are also present in the VRML2.0 specification, the 3-D sound model is also inherited from the VRML. This model provides functionalities such as positioning of sound sources in a 3-D space so that when the scene is rendered, the sounds are processed so that they appear to come from the specified location in the virtual world. The VRML also enables simple sound source directivity modeling, providing two nested ellipsoids of audibility around the sound location. The source directivity and distance dependent attenuation is implemented by decreasing the gain by which the sound samples are multiplied as a function of distance from the inner to the outer sound ellipsoid. Other specifications, such as DirectSound [9], and Java3D [10] Application Programming Interfaces (APIs), give related capabilities for presentation of 3-D positioned sound, and even simple modeling of environmental effects, such as Doppler effect and reverberation.

The second version of MPEG-4, however, offers a more versatile framework for modeling the sound propagation in an interactive virtual reality environment [11]. The part specifying these advanced sound features is called Advanced AudioBIFS, and it includes more detailed modeling of sound directivity, and sound propagation from the source to the listener taking into account sound attenuation and absorption in the air, and phenomena such as sound reflections and transmission caused by interfering objects in the medium, like was discussed in Section 2. Also, Doppler shift can be automatically computed from the relative motion between the sound source and the listener, and the information about the speed of sound in the medium. Ad-
ditionally, reverberation can be added to sound by defining a frequency dependent reverberation time, thus adding realism to reverberating virtual spaces. The rendering of the acoustic environment can be defined for each sound source separately, and in specified 3-D regions in the scenes. This enables building up virtual scenes where different acoustic spaces such as rooms or halls can be present, and the sound is rendered according to the room that the user and the source itself are presently in.

4 Creating Interactive Acoustic MPEG-4 Scenes

In this section, we describe the use of the new audio scene description features in MPEG-4 version 2. Figure 1 illustrates a BIFS scene graph where Advanced AudioBIFS is taken advantage of in creating audiovisual virtual reality. An Advanced AudioBIFS node called AcousticScene gives acoustic properties to be applied within a specified 3-D region in a scene within which the sound is rendered. It is used in binding together objects that affect the sound path, and it can also be given parameters that specify reverberation that can be added to sound sources that are within the influence of the AcousticScene. The geometry nodes may be used to build up, for example, rooms whose acoustics correspond to the visual scene. For example, a visible wall may have reflectivity attached to it, and the delay and magnitude of the sound reflection off that surface depends on the current positions of the sound source and the listener. Figure 2 further shows an example of the functionality of a BIFS scene with different acoustic spaces present. In this example two AcousticScenes with different impulse responses are spatially partly separated by 3-D boxes.

Depending on where the user currently moves in the scene, he/she can hear the sound sources and the acoustics of one of the rooms. When the listener is outside of both acoustic regions, no sound is heard.

MPEG-4 can be used to build up acoustic environments in a very flexible way. The scene author can affect the level of detail with which each sound source is presented. E.g., in the same scene there may be sources which are processed according to detailed acoustic room description and frequency dependent directivity specification, and sources which can be completely non-spatial (i.e., ambient sources), or sources which positional 3-D properties and doppler effect but no wall reflections or reverberation.

5 DSP Implementation of Acoustic Scenes

Figure 3 shows an example DSP implementation of an acoustic response in a BIFS scene. In this filter structure all the parameters (delays and filter coefficients) are derived from the parametric presentation of the scene In Fig. 3, the direct sound and the reflections are taken out of a delay line, where the length of each delay depends on the relative distance between the listener and the sound source (or the image source, in the case of a sound reflection). Before the sound is processed according to the direction of its arrival at the listening point, it may be filtered according to sound transmission factors, should there be sound obstructing objects on the path of the direct sound, and reflections are filtered by the reflectivity filters that are associated to the visual walls with acoustic properties. If there is reverberation time specified for that AcousticScene, the direct sound is led through a reverberator that implements it.

In the case of an interactive audiovisual scene, there are several factors that add complexity to the processing of the sound sources. One is that when the room impulse response changes as the listener or sound source moves, or there are changes in the parameters of the acoustic environment, the DSP structure becomes time variant. The changes can occur on the coefficients of individual filters (e.g., the response of the source directivity filtering changes as the angular distance between the listener and the source changes), or new filters have to be created or deleted as for example new sound reflection becomes visible to the listening point, or

![Figure 1: An example of the node hierarchy of a scene where Advanced AudioBIFS nodes are used to add acoustic response to the environment. In this scene, the nodes which are in the shadowed box are involved in the audio processing.](image)

![Figure 2: Wire frame models of 2 rooms which both contain acoustic surfaces that reflect part of the sound and transmit part of the sound through it. Sound source S1 is within the area of AcousticScene1, and is therefore processed with the room response of that AcousticScene. Thus listener L1 hears the direct sound and the sound reflected off the ceiling and floor, the reflectivities of which are specified by the transfer functions R1(z) and H1(z), respectively. Correspondingly, the listener L3 hears the sound source S2 in AcousticScene2, and since listener L2 is inside the overlapping area of AcousticScene1 and AcousticScene2, he can hear both sound sources through the walls of the rooms, filtered by the transmission functions H3(z) and H4(z).](image)
a wall appears between the source and the listener, and in the case of the listener moving into or out from an AcousticScene, the whole DSP processing has to be instanciated or deleted, correspondingly. Also, when the distances between the listener and the sound sources change, there are corresponding changing delays in the outputs from the delay line. All the mentioned changes in the scene require the parameters to be updated in a way that the changes do not create disturbing audible effects.

As the whole visual scene is updated at certain time intervals, these updates are used for updating the acoustic responses of the currently active DSP structures as well. This incorporates an additional delay to the acoustic scene with respect to the visual scene, that is of the same length as the time between consecutive scene updates, and it should be taken care of that it is not long enough to disturb the consistency of the audio and virtual reality. In updating the DSP structure, the delays and the filter coefficients may need to be interpolated to avoid audible clicks in the response. Also, it may appear necessary to fade in and out processing modules when they are created or deleted.

As the acoustic VR processing consumes a lot of CPU time and memory, the scenes should be designed in a way that only relevant sound sources are processed at each time. Also, the sound scene should be constructed so that no needless processing is done to any of the active sound sources. Like was mentioned in Section 4, the scene designer may choose parameters so that only a necessary level of modeling is done to each sound source. Also, by separating the AcousticScene areas it can be avoided that many signal processing objects are executed at the same time.

6 Conclusions

In this paper, we have presented properties and examples of virtual acoustics rendering in the MPEG-4 multimedia standard. The upcoming second version of MPEG-4 (to be finalized in early 2000)

Figure 3: A DSP filter structure for spatial audio rendering.

will feature a powerful set of tools for natural and synthetic audio as well as a platform for post-processing, mixing, manipulation, and spatialization of audio alone or in relation to visual rendering. Prospective application areas for MPEG-4 virtual audio technology are in multimedia and virtual reality in general, and specifically related to computer music in audiovisual installations, interactive music, and simulation technologies.

7 Acknowledgments

This work has been financially supported by Nokia Research Center and the Pythagoras Graduate School.

References


