INTERNATIONAL STANDARD

ISO/IEC 14496-11

First edition
2005-12-15

Information technology — Coding of audio-visual objects —

Part 11: Scene description and application engine

Technologies de l'information — Codage des objets audiovisuels —
Partie 11: Description de scène et moteur d'application

Reference number
ISO/IEC 14496-11:2005(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

ISO/IEC 14496-11 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

ISO/IEC 14496 consists of the following parts, under the general title *Information technology — Coding of audio-visual objects*:

— *Part 1: Systems*

— *Part 2: Visual*

— *Part 3: Audio*

— *Part 4: Conformance testing*

— *Part 5: Reference software*

— *Part 6: Delivery Multimedia Integration Framework (DMIF)*

— *Part 7: Optimized reference software for coding of audio-visual objects* [Technical Report]

— *Part 8: Carriage of ISO/IEC 14496 contents over IP networks*

— *Part 9: Reference hardware description* [Technical Report]

— *Part 10: Advanced Video Coding*

— *Part 11: Scene description and application engine*

— *Part 12: ISO base media file format*

— *Part 13: Intellectual Property Management and Protection (IPMP) extensions*

— *Part 14: MP4 file format*

— *Part 15: Advanced Video Coding (AVC) file format*

— *Part 16: Animation Framework eXtension (AFX)*

— *Part 17: Streaming text format*

— *Part 18: Font compression and streaming*

— *Part 19: Synthesized texture stream*

— *Part 20: Lightweight Application Scene Representation (LASeR) and Simple Aggregation Format (SAF)*

The following parts are under preparation:

— *Part 21: MPEG-J GFX*

0 Introduction

0.1 Scene Description

0.1.1 Overview

ISO/IEC 14496 addresses the coding of audio-visual objects of various types: natural video and audio objects as well as textures, text, 2- and 3-dimensional graphics, and also synthetic music and sound effects. To reconstruct a multimedia scene at the terminal, it is hence not sufficient to transmit the raw audio-visual data to a receiving terminal. Additional information is needed in order to combine this audio-visual data at the terminal and construct and present to the end-user a meaningful multimedia scene. This information, called scene description, determines the placement of audio-visual objects in space and time and is transmitted together with the coded objects as illustrated in Figure 1. Note that the scene description only describes the structure of the scene. The action of assembling these objects in the same representation space is called composition. The action of transforming these audio-visual objects from a common representation space to a specific presentation device (i.e. speakers and a viewing window) is called rendering.

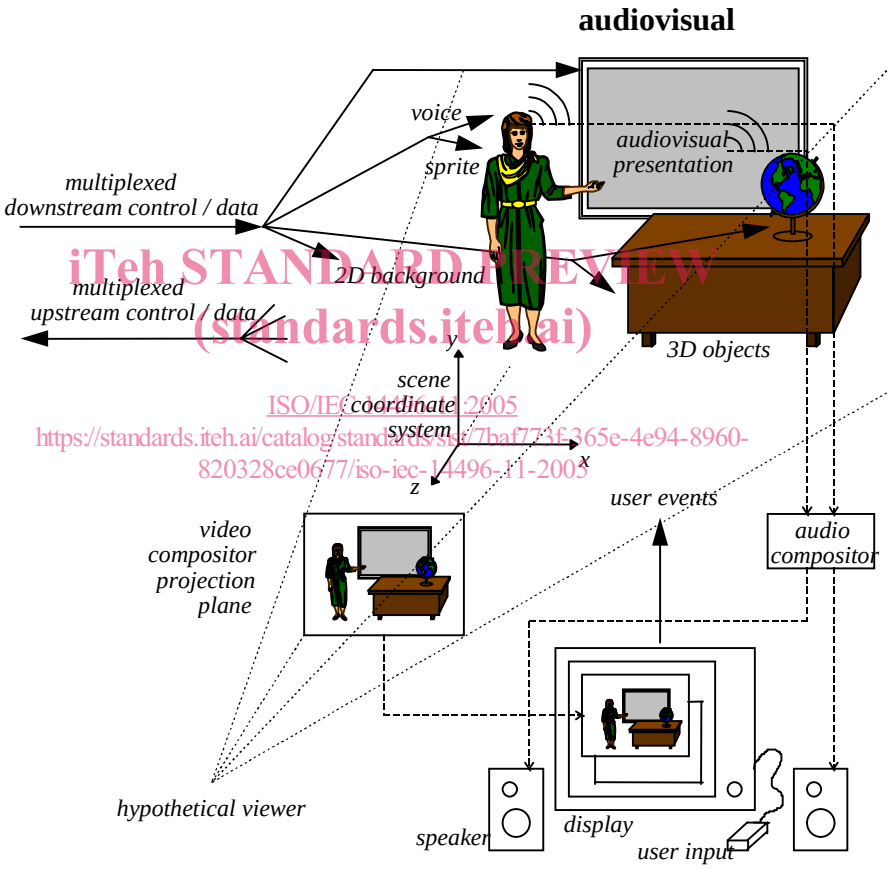

Figure 1 — An example of an object-based multimedia scene

Independent coding of different objects may achieve higher compression, and also brings the ability to manipulate content at the terminal. The behaviors of objects and their response to user inputs can thus also be represented in the scene description.

The scene description framework used in this part of ISO/IEC 14496 is based largely on ISO/IEC 14772-1:1998 (Virtual Reality Modeling Language – VRML).

0.1.2 Composition and Rendering

ISO/IEC 14496-11 defines the syntax and semantics of bitstreams that describe the spatio-temporal relationships of audio-visual objects. For visual data, particular composition algorithms are not mandated since they are implementation-

dependent; for audio data, subclause 7.1.1.2.13 and the semantics of the AudioBIFS nodes normatively define the composition process. The manner in which the composed scene is presented to the user is not specified for audio or visual data. The scene description representation is termed “BInary Format for Scenes” (BIFS).

0.1.3 Scene Description

In order to facilitate the development of authoring, editing and interaction tools, scene descriptions are coded independently from the audio-visual media that form part of the scene. This permits modification of the scene without having to decode or process in any way the audio-visual media. The following clauses detail the scene description capabilities that are provided by ISO/IEC 14496-11.

0.1.3.1 Grouping of audio-visual objects

A scene description follows a hierarchical structure that can be represented as a graph. Nodes of the graph form audio-visual objects, as illustrated in Figure 2. The structure is not necessarily static; nodes may be added, deleted or be modified.

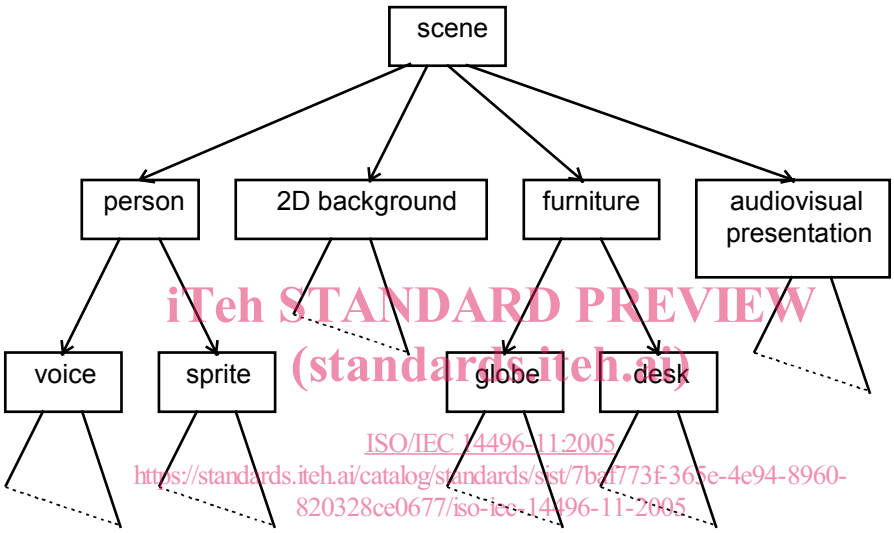

Figure 2 — Logical structure of example scene

0.1.3.2 Spatio-Temporal positioning of objects

Audio-visual objects have both a spatial and a temporal extent. Complex audio-visual objects are constructed by combining appropriate scene description nodes to build up the scene graph. Audio-visual objects may be located in 2D or 3D space. Each audio-visual object has a local co-ordinate system. A local co-ordinate system is one in which the audio-visual object has a pre-defined (but possibly varying) spatio-temporal location and scale (size and orientation). Audio-visual objects are positioned in a scene by specifying a co-ordinate transformation from the object’s local co-ordinate system into another co-ordinate system defined by a parent node in the scene graph.

0.1.3.3 Attributes of audio-visual objects

Scene description nodes expose a set of parameters through which aspects of their appearance and behavior can be controlled.

EXAMPLE — the volume of a sound; the color of a synthetic visual object; the source of a streaming video.

0.1.3.4 Behavior of audio-visual objects

ISO/IEC 14496-11 provides tools for enabling dynamic scene behavior and user interaction with the presented content. User interaction can be separated into two major categories: client-side and server-side. Client-side interaction is an integral part of the scene description described herein. Server-side interaction is not dealt with.

Client-side interaction involves content manipulation that is handled locally at the end-user’s terminal. It consists of the modification of attributes of scene objects according to specified user actions.

EXAMPLE — A user can click on a scene to start an animation or video sequence. The facilities for describing such interactive behavior are part of the scene description, thus ensuring the same behavior in all terminals conforming to ISO/IEC 14496-11.

0.2 Extensible MPEG-4 Textual Format

0.2.1 Overview

The Extensible MPEG-4 Textual format (XMT) is a framework (illustrated in Figure 3) for representing MPEG-4 scene description using a textual syntax. The XMT allows the content authors to exchange their content with other authors, tools or service providers, and facilitates interoperability with both the Extensible 3D (X3D) being developed by the Web3D and the Synchronized Multimedia Integration Language (SMIL) from the W3C.

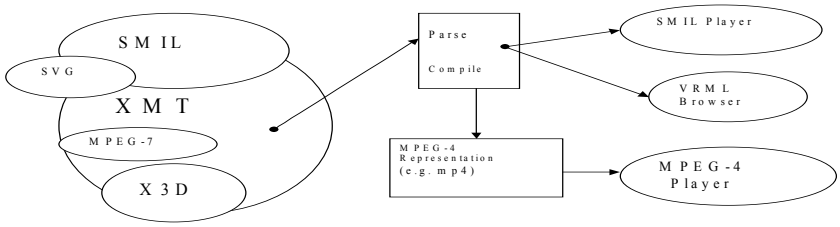

Figure 3 — Overview of the XMT Framework

0.2.2 Interoperability of XMT

The XMT format can be interchangeable between SMIL players, VRML players, and MPEG-4 players. The format can be parsed and played directly by a W3C SMIL player, preprocessed to Web3D X3D and played back by a VRML player, or compiled to an MPEG-4 representation such as MP4, which can then be played by an MPEG-4 player. See below for a graphical description of interoperability of the XMT.

0.2.3 Two-tier Architecture: XMT-A and XMT-Ω Formats

The XMT framework consists of two levels of textual syntax and semantics: the XMT-A format and the XMT-Ω format, which we will abbreviate by A and Ω, respectively, and use them interchangeably where there is no confusion.

The XMT-A is an XML-based version of MPEG-4 content, which contains a subset of the X3D. Also contained in XMT-A is an MPEG-4 extension to the X3D to represent MPEG-4 specific features. The XMT-A provides a straightforward, one-to-one mapping between the textual and binary formats.

The XMT-Ω is a high-level abstraction of MPEG-4 features designed based on the W3C SMIL. The XMT provides a default mapping from Ω to A, for there is no deterministic mapping between the two, and it also provides content authors with an escape mechanism from Ω to A.

In addition an XMT-C (Common) section contains the definition of elements and attributes that may be used within either XMT-A or XMT-Ω.

0.3 MPEG-J

0.3.1 Overview

MPEG-J is a flexible programmatic control system that represents an audio-visual session in a manner that allows the session to adapt to the operating characteristics when presented at the terminal. Two important characteristics are supported: first, the capability to allow graceful degradation under limited or time varying resources, and second, the ability to respond to user interaction and provide enhanced multimedia functionality.

More specifically, 9.7 normatively defines:

The format and delivery of Java byte code by specifying the MPEG-J stream format and the delivery mechanism of such a stream (Java byte code and associated data);

The MPEG-J Session and the MPEG-J application lifecycle; and

The interactions and behavior of byte code through the specification of Java APIs.

0.3.2 Organization MPEG-J specification

10.1 gives an overall architecture of the MPEG-J system. MPEG-J Session start-up is walked through in 10.2. The Delivery of MPEG-J data to the terminal is specified in 10.3. 10.4 specifies the different categories of APIs that a program in the form of Java bytecode would use. 10.5 is an informative annex on starting the Java Virtual Machine. The electronic annex attached to this document lists the normative MPEG-J APIs in the HTML format. 10.6 illustrates the usage of MPEG-J APIs through a few examples.

The International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) draw attention to the fact that it is claimed that compliance with this document may involve the use of patents.

The ISO and IEC take no position concerning the evidence, validity and scope of these patent rights.

The holder of these patent rights have assured the ISO and IEC that they are willing to negotiate licences under reasonable and non-discriminatory terms and conditions with applicants throughout the world. In this respect, the statement of the holder of this patent right is registered with the ISO and IEC. Information may be obtained from the companies listed in Annex D.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights other than those identified in Annex D. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Information technology — Coding of audio-visual objects —

Part 11:
Scene description and application engine

1 Scope

This part of ISO/IEC 14496 specifies:

1. the coded representation of the spatio-temporal positioning of audio-visual objects as well as their behavior in response to interaction (scene description);
2. the Extensible MPEG-4 Textual (XMT) format, a textual representation of the multimedia content described in ISO/IEC 14496 using the Extensible Markup Language (XML); and
3. a system level description of an application engine (format, delivery, lifecycle, and behavior of downloadable Java byte code applications).

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 639-2:1998, *Codes for the representation of names of languages — Part 2: Alpha-3 code*

ISO 3166-1:1997, *Codes for the representation of names of countries and their subdivisions — Part 1: Country codes*

ISO 9613-1:1993, *Acoustics — Attenuation of sound during propagation outdoors — Part 1: Calculation of the absorption of sound by the atmosphere*

ISO/IEC 11172-2:1993, *Information technology — Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s — Part 2: Video*

ISO/IEC 11172-3:1993, *Information technology — Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s — Part 3: Audio*

ISO/IEC 13818-3:1998, *Information technology — Generic coding of moving pictures and associated audio information — Part 3: Audio*

ISO/IEC 13818-7: 2004, *Information technology — Generic coding of moving pictures and associated audio information — Part 7: Advanced Audio Coding (AAC)*

ISO/IEC 14496-2:2004, *Information technology — Coding of audio-visual objects — Part 2: Visual*

ISO/IEC 14772-1:1997, *Information technology — Computer graphics and image processing — The Virtual Reality Modeling Language — Part 1: Functional specification and UTF-8 encoding*

ISO/IEC 14772-1:1997/Amd.1:2003, *Information technology — Computer graphics and image processing — The Virtual Reality Modeling Language — Part 1: Functional specification and UTF-8 encoding — Amendment 1: Enhanced interoperability*

ISO/IEC 16262:2002, *Information technology — ECMAScript language specification*

ISO/IEC 13818-2:2000, *Information technology — Generic coding of moving pictures and associated audio information — Part 2: Video*

ISO/IEC 10918-1:1994, *Information technology — Digital compression and coding of continuous-tone still images: Requirements and guidelines*

IEEE Std 754-1985, *Standard for Binary Floating-Point Arithmetic*

Addison-Wesley:September 1996, *The Java Language Specification,* by James Gosling, Bill Joy and Guy Steele, ISBN 0-201-63451-1

Addison-Wesley:September 1996, *The Java Virtual Machine Specification,* by T. Lindholm and F. Yellin, ISBN 0-201-63452-X

Addison-Wesley:July 1998, *Java Class Libraries Vol. 1 The Java Class Libraries,* Second Edition Volume 1, by Patrick Chan, Rosanna Lee and Douglas Kramer, ISBN 0-201-31002-3

Addison-Wesley:July 1998, *Java Class Libraries Vol. 2 The Java Class Libraries,* Second Edition Volume 2, by Patrick Chan and Rosanna Lee, ISBN 0-201-31003-1

Addison-Wesley, May 1996, *Java API, The Java Application Programming Interface, Volume 1: Core Packages,* by J. Gosling, F. Yellin and the Java Team, ISBN 0-201-63453-8

DAVIC 1.4.1 specification Part 9: Information Representation

ANSI/SMPTE 291M-1996, *Television — Ancillary Data Packet and Space Formattin*g

SMPTE 315M -1999, *Television — Camera Positioning Information Conveyed by Ancillary Data Packets*

3 Additional reference

ISO/IEC 13522-6:1998, *Information technology — Coding of multimedia and hypermedia information — Part 6: Support for enhanced interactive applications.* This reference contains the full normative references to Java APIs and the Java Virtual Machine as described in the normative references above.

4 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

4.1
Access Unit (AU)

individually accessible portion of data within an *elementary stream*

NOTE An access unit is the smallest data entity to which timing information can be attributed.

4.2
Alpha map

representation of the transparency parameters associated with a texture map

4.3
atom

object-oriented building block defined by a unique type identifier and length

4.4
audio-visual object

representation of a natural or synthetic object that has an audio and/or visual manifestation

NOTE The representation corresponds to a node or a group of nodes in the BIFS scene description. Each audio-visual object is associated with zero or more *elementary streams* using one or more *object descriptors.*

4.5
audio-visual scene (AV scene)

set of audio-visual objects together with scene description information that defines their spatial and temporal attributes including behaviors resulting from object and user interactions

4.6
Binary Format for Scene (BIFS)

coded representation of a parametric scene description format

4.7
buffer model

model that defines how a terminal complying with ISO/IEC 14496 manages the buffer resources that are needed to decode a presentation

4.8
byte aligned
position in a coded bit stream with a distance of a multiple of 8-bits from the first bit in the stream

4.9
chunk
contiguous set of samples stored for one stream

4.10
clock reference
special time stamp that conveys a reading of a time base

4.11
composition
process of applying scene description information in order to identify the spatio-temporal attributes and hierarchies of audio-visual objects

4.12
Composition Memory (CM)
random access memory that contains composition units

4.13
Composition Time Stamp (CTS)
indication of the nominal composition time of a composition unit

4.14
Composition Unit (CU)
individually accessible portion of the output that a decoder produces from access units

4.15
compression layer
layer of a system according to the specifications in ISO/IEC 14496 that translates between the coded representation of an elementary stream and its decoded representation

NOTE It incorporates the decoders.

4.16
container atom
atom whose sole purpose is to contain and group a set of related atoms

4.17
decoder
entity that translates between the coded representation of an elementary stream and its decoded representation

4.18
decoding buffer (DB)
buffer at the input of a decoder that contains access units

4.19
decoder configuration
configuration of a decoder for processing its elementary stream data by using information contained in its elementary stream descriptor

4.20
Decoding Time Stamp (DTS)
indication of the nominal decoding time of an access unit

4.21
delivery layer
generic abstraction for delivery mechanisms (computer networks, etc.) able to store or transmit a number of multiplexed elementary streams or M4Mux streams

4.22
descriptor
data structure that is used to describe particular aspects of an elementary stream or a coded audio-visual object

4.23
DMIF Application Interface (DAI)
interface specified in ISO/IEC 14496-6

NOTE It is used here to model the exchange of SL-packetized stream data and associated control information between the sync layer and the delivery layer.

4.24
Elementary Stream (ES)

consecutive flow of mono-media data from a single source entity to a single destination entity on the compression layer

4.25
Elementary Stream Descriptor (ESD)

structure contained in object descriptors that describes the encoding format, initialization information, sync layer configuration, and other descriptive information about the content carried in an elementary stream

4.26
Elementary Stream Interface (ESI)

conceptual interface modeling the exchange of elementary stream data and associated control information between the compression layer and the sync layer

4.27
M4Mux Channel (FMC)

label to differentiate between data belonging to different constituent streams within one M4Mux Stream

NOTE A sequence of data in one M4Mux channel within a M4Mux stream corresponds to one single SL-packetized stream.

4.28
M4Mux packet

smallest data entity managed by the M4Mux tool, consisting of a header and a payload

4.29
M4Mux stream

sequence of M4Mux Packets with data from one or more SL-packetized streams that are each identified by their own M4Mux channel

4.30
M4Mux tool

tool that allows the interleaving of data from multiple data streams

4.31
graphics profile

profile that specifies the permissible set of graphical elements of the BIFS tool that may be used in a scene description stream

NOTE Note BIFS comprises both graphical and scene description elements.

4.32
hint track

special track which contains instructions for packaging one or more tracks into a TransMux

NOTE It does not contain media data (an elementary stream).

4.33
hinter

tool that is run on a completed file to add one or more hint tracks to the file to facilitate streaming

4.34
inter

mode for coding parameters that uses previously coded parameters to construct a prediction

4.35
interaction stream

elementary stream that conveys user interaction information

4.36
intra

mode for coding parameters that does not make reference to previously coded parameters to perform the encoding

4.37
initial object descriptor

special object descriptor that allows the receiving terminal to gain initial access to portions of content encoded according to ISO/IEC 14496

NOTE It conveys profile and level information to describe the complexity of the content.

4.38
Intellectual Property Identification (IPI)
unique identification of one or more elementary streams corresponding to parts of one or more audio-visual objects

4.39
Intellectual Property Management and Protection System (IPMP)
generic term for mechanisms and tools to manage and protect intellectual property

NOTE Only the interface to such systems is normatively defined.

4.40
media node
following list of time dependent nodes that refers to a media stream through a URL field: **AnimationStream**, **AudioBuffer, AudioClip**, **AudioSource**, **Inline**, **MovieTexture**

4.41
media stream
one or more elementary streams whose ES descriptors are aggregated in one object descriptor and that are jointly decoded to form a representation of an AV object

4.42
media time line
time line expressing normal play back time of a media stream

4.43
movie atom
container atom whose sub-atoms define the meta-data for a presentation ('moov')

4.44
movie data atom
container atom which can hold the actual media data for a presentation ('mdat')

4.45
MP4 file
name of the file format described in this specification

4.46
Object Clock Reference (OCR)
clock reference that is used by a decoder to recover the time base of the encoder of an elementary stream

4.47
Object Content Information (OCI)
additional information about content conveyed through one or more elementary streams

NOTE It is either aggregated to individual elementary stream descriptors or is itself conveyed as an elementary stream.

4.48
Object Descriptor (OD)
descriptor that aggregates one or more elementary streams by means of their elementary stream descriptors and defines their logical dependencies

4.49
Object Descriptor Command
command that identifies the action to be taken on a list of object descriptors or object descriptor IDs, e.g. update or remove

4.50
Object Descriptor Profile
profile that specifies the configurations of the object descriptor tool and the sync layer tool that are allowed

4.51
Object Descriptor Stream
elementary stream that conveys object descriptors encapsulated in object descriptor commands

4.52
Object Time Base (OTB)
time base valid for a given elementary stream, and hence for its decoder

NOTE The OTB is conveyed to the decoder via object clock references. All time stamps relating to this object's decoding process refer to this time base.

4.53
Parametric Audio Decoder
set of tools for representing and decoding speech signals coded at bit rates between 6 Kbps and 16 Kbps, according to the specifications in ISO/IEC 14496-3