INTERNATIONAL STANDARD

ISO/IEC 14496-3

First edition
1999-12-15

Technologies de l'information — Codage des objets audiovisuels —

Partie 3: Codage audio

Reference number
ISO/IEC 14496-3:1999(E)

Informa[redacted]echnology — Coding of audio-v objects —

Part 3:
Audio

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 734 10 79
E-mail copyright@iso.ch
Web www.iso.ch

Printed in Switzerland

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. 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.

Attention is drawn to the possibility that some of the elements of this part of ISO/IEC 14496 may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

International Standard ISO/IEC 14496-3 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 testing*
- *Part 6: Delivery Multimedia Integration Framework (DMIF)*

Annexes 2.A to 2.C, 3.C, 4.A and 5.A form a normative part of this part of ISO/IEC 14496. Annexes 1.A, 1.B, 2.D, 3.A, 3.B, 3.D to 3.F, 4.B and 5.B to 5.G are for information only.

Due to its technical nature, this part of ISO/IEC 14496 requires a special format as several standalone electronic files and, consequently, does not conform to some of the requirements of the ISO/IEC Directives, Part 3.

Information technology — Coding of audio-visual objects —

Part 3: Audio

Subpart 1: Main

Structure of this part of ISO/IEC 14496:

This part of ISO/IEC 14496 comprises six subparts:

Subpart 1:	Main
Subpart 2:	Speech coding - HVXC
Subpart 3:	Speech coding - CELP
Subpart 4:	General Audio coding (GA)
Subpart 5:	Structured audio
Subpart 6:	Text to speech interface

For reasons of manageability of large documents, this part of ISO/IEC 14496 is divided into six files, corresponding to the six subparts of the standard:

1. a025035e.pdf contains Subpart 1.
2. b025035e.pdf contains Subpart 2.
3. c025035e.pdf contains Subpart 3.
4. d025035e.pdf contains Subpart 4.
5. e025035e.pdf contains Subpart 5.
6. f025035e.pdf contains Subpart 6.

Contents for Subpart 1

Subpart 1: Main

1.1 Scope

1.1.1 Overview of MPEG-4 Audio

This part of ISO/IEC 14496 (MPEG-4 Audio) is a new kind of audio standard that integrates many different types of audio coding: natural sound with synthetic sound, low bitrate delivery with high-quality delivery, speech with music, complex soundtracks with simple ones, and traditional content with interactive and virtual-reality content. By standardizing individually sophisticated coding tools as well as a novel, flexible framework for audio synchronization, mixing, and downloaded post-production, the developers of the MPEG-4 Audio standard have created new technology for a new, interactive world of digital audio.

MPEG-4, unlike previous audio standards created by ISO/IEC and other groups, does not target a single application such as real-time telephony or high-quality audio compression.Rather, MPEG-4 Audio is a standard that applies to *every* application requiring the use of advanced sound compression, synthesis, manipulation, or playback. The subparts that follow specify the state-of-the-art coding tools in several domains; however, MPEG-4 Audio is more than just the sum of its parts. As the tools described here are integrated with the rest of the MPEG-4 standard, exciting new possibilities for object-based audio coding, interactive presentation, dynamic soundtracks, and other sorts of new media, are enabled.

Since a single set of tools is used to cover the needs of a broad range of applications, *interoperability* is a natural feature of systems that depend on the MPEG-4 Audio standard. A system that uses a particular coder—for example, a real-time voice communication system making use of the MPEG-4 speech coding toolset—can easily share data and development tools with other systems, even in different domains, that use the same tool—for example, a voicemail indexing and retrieval system making use of MPEG-4 speech coding. A multimedia terminal that can decode the Main Profile of MPEG-4 Audio has audio capabilities that cover the entire spectrum of audio functionality available today and in the future.

The remainder of this clause gives a more detailed overview of the capabilities and functioning of MPEG-4 Audio. First, a discussion of concepts that have changed since the MPEG-2 audio standards is presented; then, the MPEG-4 Audio toolset is outlined.

1.1.2 New concepts in MPEG-4 Audio

Many concepts in MPEG-4 Audio are different than those in previous MPEG Audio standards. For the benefit of readers who are familiar with MPEG-1, MPEG-2, and MPEG-AAC, we provide a brief overview here.

- **MPEG-4 has no standard for transport.** In all of the MPEG-4 tools for audio and visual coding, the coding standard ends at the point of constructing a sequence of *access units*

  interface (the Delivery Multimedia Interface Format, or DMIF, specified in ISO/IEC 14496-6) that describes the capabilities of a transport layer and the communication between transport, multiplex, and demultiplex functions in encoders and decoders. The use of DMIF and the MPEG-4 Systems bitstream specification allows transmission functions that are much more sophisticated than are possible with previous MPEG standards.

  For applications which do not require sophisticated transport functionality, object-based coding, synchronization with other media, or other functions provided by MPEG-4 Systems, a private, not normative transport may be used to deliver a single MPEG-4 Audio stream. An example private transport for this purpose is given in Informative Annex A of subpart 1.

- **MPEG-4 Audio encourages low-bitrate coding.** Previous MPEG Audio standards have focused primarily on transparent (undetectable) or nearly transparent coding of high-quality audio at whatever bitrate was required to provide it. MPEG-4 provides new and improved tools for this purpose, but also standardizes (and has tested) tools that can be used for transmitting audio at the low bitrates suitable for Internet, digital radio, or other bandwidth-limited delivery. The tools specified in MPEG-4 are the state-of-the-art tools available for low-bitrate coding of speech and other audio.

- **MPEG-4 is an object-based coding standard with multiple tools.** Previous MPEG Audio standards provided a single toolset, with different configurations of that toolset specified for use in various applications. MPEG-4 provides several toolsets that have no particular relationship to each other, each with a different target function. The Profiles of MPEG-4 Audio (subclause 5.1) specify which of these tools are used together for various applications.

  Further, in previous MPEG standards, a single (perhaps multi-channel or multi-language) piece of content was all that was transmitted. In MPEG-4, by contrast, the concept of a *soundtrack* is much more flexible. Multiple tools may be used to transmit several *audio objects*; when using multiple tools together, an *audio composition* system is used to create a single soundtrack from the audio substreams. User interaction, terminal capability, and speaker configuration may be used when determining how to produce a single soundtrack from the component objects. This capability allows significant advantages in quality and flexibility in MPEG-4 over previous audio standards.

- **MPEG-4 provides capabilities for synthetic sound.** In natural sound coding, an existing sound is compressed by a server, transmitted and decompressed at the receiver. This type of coding is the subject of many existing standards for sound compresson. MPEG-4 also standardizes a novel paradigm in which synthetic sound descriptions, including synthetic speech and synthetic music, are transmitted and then *synthesized* into sound at the receiver. Such capabilities open up new areas of very-low-bitrate but still very-high-quality coding.

As with previous MPEG standards, MPEG-4 does not standardize methods for encoding sound. Thus, content authors are left to their own decisions for the best method of creating bitstreams. At the present time, it is an open problem how to automatically convert natural sound into synthetic or multi-object descriptions; therefore, most immediate solutions will involve hand-authoring the content stream in some way. This process is similar to current schemes for MIDI-based and multi-channel mixdown authoring of soundtracks.

1.1.3 MPEG-4 Audio capabilities

1.1.3.1 Overview of capabilities

The MPEG-4 Audio tools can be broadly organized into several categories:

1. *Speech* tools for the transmission and decoding of synthetic and natural speech
2. *Audio* tools for the transmission and decoding of recorded music and other audio soundtracks
3. *Synthesis* tools for very low bitrate description and transmission, and terminal-side synthesis, of synthetic music and other sounds
4. *Composition* tools for object-based coding, interactive functionality, and audiovisual synchronization
5. *Scalability* tools for the creation of bitstreams that can be transmitted, without recoding, at several different bitrates

Each of these types of tools will be described in more detail in the following subclauses.

1.1.3.2 MPEG-4 speech coding tools

1.1.3.2.1 Introduction

Two types of speech coding tools are provided in MPEG-4. The *natural* speech tools allow the compression, transmission, and decoding of human speech, for use in telephony, personal communication, and surveillance applications. The *synthetic* speech tool provides an interface to text-to-speech synthesis systems; using synthetic speech provides very-low-bitrate operation and built-in connection with facial animation for use in low-bitrate videoteleconferencing applications. Each of these tools will be discussed.

1.1.3.2.2 Natural speech coding

The MPEG-4 speech coding toolset covers the compression and decoding of natural speech sound at bitrates ranging between 2 and 24 kbit/s. When the variable bitrate coding is allowed, coding at even less than 2 kbit/s, such as average bitrate of 1.2 kbit/s, is also supported. Two basic speech coding techniques are used: One is a parametric speech coding algorithm, HVXC (Harmonic Vector eXcitation Coding), for very low bit rates; and the other is a CELP (Code Excited Linear Prediction) coding technique. The MPEG-4 speech coder targets applications from mobile and satellite communications, to Internet telephony, to packaged media and speech databases. It meets a wide range of requirements covering bitrates, functionality and sound quality and is specified in subparts 2 and 3.

MPEG-4 HVXC operates at fixed bitrates between 2.0 kbit/s and 4.0 kbit/s, using a bitrate scalability technique.It also operates at lower bitrates, typically 1.2-1.7 kbit/s, in variable bitrate mode. HVXC provides communications-quality to near-toll-quality speech in the 100-3800 Hz band at 8kHz sampling rate. HVXC also allows independent change of speed and pitch during decoding, which is a powerful functionality for fast access to speech databases.

MPEG-4 CELP is a well-known coding algorithm with new functionality. Conventional CELP coders offer compression at a single bit rate and are optimized for specific applications. Compression is one of the functionalities provided by MPEG-4 CELP, but MPEG-4 also enables the use of one basic coder in multiple applications. It provides scalability in bitrate and bandwidth, as well as the ability to generate bitstreams at arbitrary bitrates. The MPEG-4 CELP coder supports two sampling rates, namely, 8 and 16 kHz. The associated bandwidths are 100 – 3800 Hz for 8 kHz sampling and 50 – 7000 Hz for 16 kHz sampling.

MPEG has conducted extensive verification testing in realistic listening conditions in order to prove the efficacy of the speech coding toolset.

1.1.3.2.3 Text-to-speech interface

Text-to-speech (TTS) capability is becoming a rather common media type and plays an important role in various multi-media application areas. For instance, by using TTS functionality, multimedia content with narration can be easily created without recording natural speech sound. Before MPEG-4, however, there was no way for a multimedia content provider to easily give instructions to an unknown TTS system. In MPEG-4, a single common interface for TTS systems is standardized. This interface allows speech information to be transmitted in the International Phonetic Alphabet (IPA), or in a textual (written) form of any language. It is specified in subpart 6.

The MPEG-4 TTS package, Hybrid/Multi-Level Scalable TTS Interface, can be considered as a superset of the conventional TTS framework. This extended TTS Interface can utilize prosodic information taken from natural speech in addition to input text and can thus generate much higher-quality synthetic speech. The interface and its bitstream format is strongly scalable in terms of this added information; for example, if some parameters of prosodic information are not available, a decoder can generate the missing parameters by rule. Normative algorithms for speech synthesis and text-to-phoneme translation are not specified in MPEG-4, but to meet the goal that underlies the MPEG-4 TTS Interface, a decoder should fully utilize all the provided information according to the user's requirements level.

As well as an interface to Text-to-speech synthesis systems, MPEG-4 specifies a joint coding method for phonemic information and facial animation (FA) parameters and other animation parameters (AP). Using this technique, a single bitstream may be used to control both the Text-to-Speech Interface and the Facial Animation visual object decoder (see ISO/IEC 14496-2 Annex C). The functionality of this extended TTS thus ranges from conventional TTS to natural speech coding and its application areas, from simple TTS to audio presentation with TTS and motion picture dubbing with TTS.

1.1.3.3 MPEG-4 general audio coding tools

MPEG-4 standardizes the coding of natural audio at bitrates ranging from 6 kbit/s up to several hundred kbit/s per audio channel for mono, two-channel-, and multi-channel-stereo signals. General high-quality compression is provided by the use of the MPEG-2 AAC standard (ISO/IEC 13818-7), with certain improvements, within the MPEG-4 tool set.At 64 kbit/s/channel and higher ranges, this coder has been found in verification testing under rigorous conditions to meet the criterion of “indistinguishable quality” as defined by the European Broadcasting Union.

Subpart 4 of MPEG-4 specifies the AAC tool set, in the General Audio coder. This coding technique uses a perceptual filterbank, a sophisticated masking model, noise-shaping techniques, channel coupling, and noiseless coding and bit-allocation to provide the maximum compression within the constraints of providing the highest possible quality. Psychoacoustic coding standards developed by MPEG have represented the state-of-the-art in this technology for nearly 10 years; MPEG-4 General Audio coding continues this tradition.

For bitrates from 6 kbit/s up to 64 kbit/s per channel, the MPEG-4 standard provides extensions to AAC and the TwinVQ tools that allow the content author to achieve highest quality by altering the tool used depending on the bit rate. Furthermore, various bit rate scalability options are available within the GA coder (see subclause 1.1.3.6.). The low-bitrate techniques and scalability modes provided with this tool set have also been verified in formal tests by MPEG.

1.1.3.4 MPEG-4 Audio synthesis tools

The MPEG-4 toolset providing general audio synthesis capability is called MPEG-4 Structured Audio, and it is described in subpart 5 of ISO/IEC 14496-3. (There is also a tool for the transmission of synthetic speech; it is described above in subclause 1.2.2 and in subpart 6). MPEG-4 Structured Audio (the SA coder) provides very general capabilities for the description of synthetic sound, and the normative creation of synthetic sound in the decoding terminal. High-quality stereo sound can be transmitted at bitrates from 0 kbit/s (no continuous cost) to 2-3 kbit/s for extremely expressive sound using these tools.

Rather than specify a particular method of synthesis, SA specifies a flexible language for describing methods of synthesis.This technique allows content authors two advantages. First, the set of synthesis techniques available is not limited to those that were envisioned as useful by the creators of the standard; any current or future method of synthesis may be used in MPEG-4 Structured Audio. Second, the creation of synthetic sound from structured descriptions is normative in MPEG-4, so sound created with the SA coder will sound the same on any terminal.

Synthetic audio is transmitted via a set of *instrument* modules that can create audio signals under the control of a *score*. An instrument is a small network of signal-processing primitives that control the parametric generation of sound according to some algorithm. Several different instruments may be transmitted and used in a single Structured Audio bitstream. A score is a time-sequenced set of commands that invokes various instruments at specific times to contribute their output to an overall music performance. The format for the description of instruments—SAOL, the Structured Audio Orchestra Language—and that for the description of scores—SASL, the Structured Audio Score Language—are specified in subpart 6.

Efficient transmission of sound samples, also called *wavetables*, for use in sampling synthesis is accomplished by providing interoperability with the MIDI Manufacturers Association Downloaded Sounds Level 2 (DLS-2) standard, which is normatively referenced by the Structured Audio standard. By using the DLS-2 format, the simple and popular technique of wavetable synthesis can be used in MPEG-4 Structured Audio soundtracks, either by itself or in conjunction with other kinds of synthesis using the more general-purpose tools. To further enable interoperability with existing content and authoring tools, the popular MIDI (Musical Instrument Digital Interface) control format can be used instead of, or in addition to, scores in SASL for controlling synthesis.

Through the inclusion of compatibility with MIDI standards, MPEG-4 Structured Audio thus represents a unification of the current technique for synthetic sound description (MIDI-based wavetable synthesis) with that of the future (general-purpose algorithmic synthesis). The resulting standard solves problems not only in very-low-bitrate coding, but also in virtual environments, video games, interactive music, karaoke systems, and many other applications.

1.1.3.5 MPEG-4 Audio composition tools

The tools for audio composition, like those for visual composition, are specified in the MPEG-4 Systems standard (ISO/IEC 14496-1). However, since readers interested in audio functionality are likely to look here first, a brief overview is provided.

Audio composition is the use of multiple individual “audio objects” and mixing techniques to create a single soundtrack.It is analogous to the process of recording a soundtrack in a multichannel mix, with each musical instrument, voice actor, and sound effect on its own channel, and then “mixing down” the multiple channels to a single channel or single stereo pair. In MPEG-4, the multichannel mix itself may be transmitted, with each audio source using a different coding tool, and a set of instructions for mixdown also transmitted in the bitstream. As the multiple audio objects are received, they are decoded separately, but not played back to the listener; rather, the instructions for mixdown are used to prepare a single soundtrack from the “raw material” given in the objects. This final soundtrack is then played for the listener.

An example serves to illustrate the efficacy of this approach. Suppose, for a certain application, we wish to transmit the sound of a person speaking in a reverberant environment over stereo background music, at very high quality. A traditional approach to coding would demand the use of a general audio coding at 32 kbit/s/channel or above; the sound source is too complex to be well-modeled by a simple model-based coder. However, in MPEG-4 we can represent the soundtrack as the conjunction of several objects: a **speaking person** passed through a **reverberator** added to a **synthetic music track**. We transmit the speaker’s voice using the CELP tool at 16 kbit/s, the synthetic music using the SA tool at 2 kbit/s, and allow a small amount of overhead (only a few hundreds of bytes as a fixed cost) to describe the stereo mixdown and the reverberation. Using MPEG-4 and an object-based approach thus allows us to describe in less than 20 kbit/s total a bitstream that might require 64 kbit/s to transmit with traditional coding, at equivalent quality.

Additionally, having such structured soundtrack information present in the decoding terminal allows more sophisticated client-side interaction to be included. For example, the listener can be allowed (if the content author desires) to request that the background music be muted. This functionality would not be possible if the music and speech were coded into the same audio track.

With the MPEG-4 Binary Format for Scenes (BIFS), specified in MPEG-4 Systems, a subset tool called AudioBIFS allows content authors to describe sound scenes using this object-based framework. Multiple sources may be mixed and combined, and interactive control provided for their combination. Sample-resolution control over mixing is provided in this method. Dynamic download of custom signal-processing routines allows the content author to exactly request a particular, normative, digital filter, reverberator, or other effects-processing routine. Finally, an interface to terminal-dependent methods of 3-D audio spatialisation is provided for the description of virtual-reality and other 3-D sound material.

As AudioBIFS is part of the general BIFS specification, the same framework is used to synchronize audio and video, audio and computer graphics, or audio with other material. Please refer to ISO/IEC 14496-1 (MPEG-4 Systems) for more information on AudioBIFS and other topics in audiovisual synchronization.

1.1.3.6 MPEG-4 Audio scalability tools

Many of the bitstream types in MPEG-4 are *scalable* in one manner or another. Several types of scalability in the standard are discussed below.

Bitrate scalability allows a bitstream to be parsed into a bitstream of lower bitrate such that the combination can still be decoded into a meaningful signal. The bitstream parsing can occur either during transmission or in the decoder. Scalability is available within each of the natural audio coding schemes, or by a combination of different natural audio coding schemes.

Bandwidth scalability is a particular case of bitrate scalability, whereby part of a bitstream representing a part of the frequency spectrum can be discarded during transmission or decoding. This is available for the CELP speech coder, where an extension layer converts the narrow band base layer encoder into a wide band speech coder. Also the general audio coding tools which all operate in the frequency domain offer a very flexible bandwidth control for the different coding layers.

Encoder complexity scalability allows encoders of different complexity to generate valid and meaningful bitstreams. An example for this is the availability of a high quality and a low complexity excitation module for the wideband CELP coder allowing to choose between significant lower encoder complexity or optimized coding quality.

Decoder complexity scalability allows a given bitstream to be decoded by decoders of different levels of complexity. A subtype of decoder complexity scalability is *graceful degradation*, in which a decoder dynamically monitors the resources available, and scales down the decoding complexity (and thus the audio quality) when resources are limited. The Structured Audio decoder allows this type of scalability; a content author may provide (for example) several different algorithms for the synthesis of piano sounds, and the content itself decides, depending on available resources, which one to use.

1.2 Normative references

The following normative documents contain provisions which, through reference in this text, constitute provisions of this part of ISO/IEC 14496. For dated references, subsequent amendments to, or revisions of, any of these publications do not apply. However, parties to agreements based on this part of ISO 14496 are encouraged to investigate the possibility of applying the most recent editions of the normative documents indicated below. For undated references, the latest edition of the normative document referred to applies. Members of ISO and IEC maintain registers of currently valid International Standards.

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*.
ITU-T Rec. H.222.0(1995) | ISO/IEC 13818-1:1996, *Information technology - Generic coding of moving pictures and associated audio information: Systems.*
ISO/IEC 13818-3:1998, *Information technology - Generic coding of moving pictures and associated audio information - Part 3: Audio*.
ISO/IEC 13818-7:1997, *Information technology - Generic coding of moving pictures and associated audio information - Part 7: Advanced Audio Coding (AAC).*
(c) 1996 MIDI Manufacturers Association, *The Complete MIDI 1.0 Detailed Specification* v. 96.2.
(c) 1998 MIDI Manufacturers Association, *The MIDI Downloadable Sounds Specification, v. 98.2*.

1.3 Terms and definitions

For the purposes of this part of ISO 14496, the following terms and definitions apply.