INTERNATIONAL STANDARD

ISO/IEC 14496-3

First edition
1999-12-15

AMENDMENT 1
2000-09-15

Information technology — Coding of audio-visual objects —

Part 3:
Audio

AMENDMENT 1: Audio extensions

Technologies de l'information — Codage des objets audiovisuels —

Partie 3: Codage audio

AMENDEMENT 1: Extensions audio

Reference number
ISO/IEC 14496-3:1999/Amd.1:2000(E)

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.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 3.

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 Amendment may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Amendment 1 to International Standard ISO/IEC 14496-3:2000 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

Annexes A, B, C, D, E and F of this Amendment are for information only.

Introduction

MPEG-4 version 2 is an amendment to MPEG-4 version 1. This document contains the description of bitstream and decoder extensions related to new tools defined within MPEG-4 version 2. As long as nothing else is mentioned, the description made in MPEG-4 version 1 is not changed but only extended.

Overview

ISO/IEC 14496-3 (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.

The following subclauses give a more detailed overview of the capabilities and functionalities provided with MPEG-4 Audio version 2.

New concepts

With this extension, new tools are added to the MPEG-4 standard, while none of the existing tools of version 1 is replaced. Version 2 is therefore fully backward compatible to version 1.

In the area of Audio, new tools are added in MPEG-4 version 2 to provide the following new functionalities:

- Error Robustness

The error robustness tools provide improved performance on error-prone transmission channels. They can be distinguished into codec specific error resilience tools and an common error protection tool.

Improved error robustness for AAC is provided by a set of error resilience tools. These tools reduce the perceived deterioration of the decoded audio signal that is caused by corrupted bits in the bitstream. The following tools are provided to improve the error robustness for several parts of an AAC frame:

- Virtual CodeBook tool (VCB11)
- Reversible Variable Length Coding tool (RVLC)
- Huffman Codeword Reordering tool (HCR)

Improved error robustness capabilities for all coding tools are provided through the error resilient bitstream payload syntax. It allows advanced channel coding techniques, which can be adapted to the special needs of the different coding tools. This error resilient bitstream payload syntax is mandatory for all version 2 object types.

The error protection tool (EP tool) provides unequal error protection (UEP) for MPEG-4 Audio in conjunction with the error resilient bitstream payload. UEP is an efficient method to improve the error robustness of source coding schemes. It is used by various speech and audio coding systems operating over error-prone channels such as mobile telephone networks or Digital Audio Broadcasting (DAB). The bits of the coded signal representation are first grouped into different classes according to their error sensitivity. Then error protection is individually applied to the different classes, giving better protection to more sensitive bits.

- Low-Delay Audio Coding

The MPEG-4 General Audio Coder provides very efficient coding of general audio signals at low bitrates. However it has an algorithmic delay of up to several 100ms and is thus not well suited for applications requiring low coding delay, such as real-time bi-directional communication. As an example, for the General Audio Coder operating at 24 kHz sampling rate and 24 kbit/s this results in an algorithmic coding delay of about 110 ms plus up to additional 210 ms for the bit reservoir. To enable coding of general audio signals with an algorithmic delay not exceeding 20 ms, MPEG-4 version 2 specifies a Low-Delay Audio Coder which is derived from MPEG-2/4 Advanced Audio Coding (AAC). It operates at up to 48 kHz sampling rate and uses a frame length of 512 or 480 samples, compared to the 1024 or 960 samples used in standard MPEG-2/4 AAC. Also the size of the window used in the analysis and synthesis filterbank is reduced by a factor of 2. No block switching is used to avoid the "look-ahead'" delay due to the block switching decision. To reduce pre-echo artefacts in case of transient signals, window shape switching is provided instead. For non-transient parts of the signal a sine window is used, while a so-called low overlap window is used in case of transient signals. Use of the bit reservoir is minimized in the encoder in order to reach the desired target delay. As one extreme case, no bit reservoir is used at all.

- Fine Grain Scalability

Bitrate scalability, also known as embedded coding, is a very desirable functionality. The General Audio Coder of version 1 supports large step scalability where a base layer bitstream can be combined with one or more enhancement layer bitstreams to utilize a higher bitrate and thus obtain a better audio quality. In a typical configuration, a 24 kbit/s base layer and two 16 kbit/s enhancement layers could be used, permitting decoding at a total bitrate of 24 kbit/s (mono), 40 kbit/s (stereo), and 56 kbit/s (stereo). Due to the side information carried in each layer, small bitrate enhancement layers are not efficiently supported in version 1. To address this problem and to provide efficient small step scalability for the General Audio Coder, the Bit-Sliced Arithmetic Coding (BSAC) tool is available in version 2. This tool is used in combination with the AAC coding tools and replaces the noiseless coding of the quantized spectral data and the scalefactors. BSAC provides scalability in steps of 1 kbit/s per audio channel, i.e. 2 kbit/s steps for a stereo signal. One base layer bitstream and many small enhancement layer bitstreams are used. The base layer contains the general side information, specific side information for the first layer and the audio data of the first layer. The enhancement streams contain only the specific side information and audio data for the corresponding layer. To obtain fine step scalability, a bit-slicing scheme is applied to the quantized spectral data. First the quantized spectral values are grouped into frequency bands. Each of these groups contains the quantized spectral values in their binary representation. Then the bits of a group are processed in slices according to their significance. Thus first all most significant bits (MSB) of the quantized values in a group are processed, etc. These bit-slices are then encoded using an arithmetic coding scheme to obtain entropy coding with minimal redundancy. Various arithmetic coding models are provided to cover the different statistics of the bit-slices. The scheme used to assign the bit-slices of the different frequency bands to the enhancement layer is constructed in a special way. This ensures that, with an increasing number of enhancement layers utilized by the decoder, quantized spectral data is refined by providing more of the less significant bits. But also the bandwidth is increased by providing bit-slices of the spectral data in higher frequency bands.

- Parametric Audio Coding

The Parametric Audio Coding tools combine very low bitrate coding of general audio signals with the possibility of modifying the playback speed or pitch during decoding without the need for an effects processing unit. In combination with the speech and audio coding tools of version 1, improved overall coding efficiency is expected for applications of object based coding allowing selection and/or switching between different coding techniques.

Parametric Audio Coding uses the Harmonic and Individual Lines plus Noise (HILN) technique to code general audio signals at bitrates of 4 kbit/s and above using a parametric representation of the audio signal. The basic idea of this technique is to decompose the input signal into audio objects which are described by appropriate source models and represented by model parameters. Object models for sinusoids, harmonic tones, and noise are utilized in the HILN coder.

This approach allows to introduce a more advanced source model than just assuming a stationary signal for the duration of a frame, which motivates the spectral decomposition used e.g. in the MPEG-4 General Audio Coder. As known from speech coding, where specialized source models based on the speech generation process in the human vocal tract are applied, advanced source models can be advantageous in particular for very low bitrate coding schemes.

Due to the very low target bitrates, only the parameters for a small number of objects can be transmitted. Therefore a perception model is employed to select those objects that are most important for the perceptual quality of the signal.

In HILN, the frequency and amplitude parameters are quantized according to the “just noticeable differences” known from psychoacoustics. The spectral envelope of the noise and the harmonic tone is described using LPC modeling as known from speech coding. Correlation between the parameters of one frame and between consecutive frames is exploited by parameter prediction. The quantized parameters are finally entropy coded and multiplexed to form a bitstream.

A very interesting property of this parametric coding scheme arises from the fact that the signal is described in terms of frequency and amplitude parameters. This signal representation permits speed and pitch change functionality by simple parameter modification in the decoder. The HILN Parametric Audio Coder can be combined with MPEG-4 Parametric Speech Coder (HVXC) to form an integrated parametric coder covering a wider range of signals and bitrates. This integrated coder supports speed and pitch change. Using a speech/music classification tool in the encoder, it is possible to automatically select the HVXC for speech signals and the HILN for music signals. Such automatic HVXC/HILN switching was successfully demonstrated and the classification tool is described in the informative Annex of the version 2 standard.

- CELP Silence Compression

The silence compression tool reduces the average bitrate thanks to compression at a lower-bitrate for silence. In the encoder, a voice activity detector is used to distinguish between regions with normal speech activity and those with silence or background noise. During normal speech activity, the CELP coding as in version 1 is used. Otherwise a Silence Insertion Descriptor (SID) is transmitted at a lower bitrate. This SID enables a Comfort Noise Generator (CNG) in the decoder. The amplitude and the spectral shape of this comfort noise are specified by energy and LPC parameters in similar methods to those in a normal CELP frame. These parameters are optionally re-transmitted in the SID and thus can be updated as required.

- Extended HVXC

In the Version 1 HVXC, variable bitrate mode of 2.0 kbit/s maximum is supported as well as 2.0 and 4.0 kbit/s fixed bitrate modes. In the Version 2 Error Resilient (ER) HVXC, the variable bitrate mode of 4.0 kbit/s maximum is additionally supported. The ER HVXC therefore provides fixed bitrate modes (2.0-4.0kbit/s) and variable bitrate mode(<2.0kbit/s, <4.0kbit/s) both in a scalable and non-scalable scheme. In the variable bitrate modes, non-speech parts are detected in unvoiced signals, and a smaller number of bits is used for these non-speech parts to reduce the average bitrate. ER HVXC provides communications-quality to near-toll-quality speech in the 100-3800 Hz band at 8kHz sampling rate. When the variable bitrate mode is allowed, operation at lower average bitrate is possible. Coded speech with variable bitrate mode at typical bitrate of 1.5kbit/s average, and at typical bitrate of 3.0kbit/s average has essentially the same quality as 2.0 kbit/s fixed rate and 4.0 kbit/s fixed rate respectively. The functionality of pitch and speed change during decoding is supported for all modes. ER HVXC has the syntax with the error sensitivity classes to be used with the EP-Tool, and the error concealment functionality is supported for the use for error-prone channel like mobile communication channels. The ER HVXC speech coder targets applications from mobile and satellite communications, to Internet telephony, to packaged media and speech databases.

Capabilities

Overview of capabilities

MPEG-4 Audio version 2 provides the following new capabilities:

- error robustness (including error resilience as well as error protection)
- low delay audio coding
- backchannel
- fine granule scalability
- parametric audio
- silence compression in CELP
- extended HVXC

Those new capabilities are discussed in more detail below.

Error robustness

Error resilience tools for AAC

Several tools are provided to increase the error resilience for AAC. These tools improve the perceptual audio quality of the decoded audio signal in case of corrupted bitstreams, which may occur e. g. in the presence of noisy transmission channels.

The Virtual CodeBooks tool (VCB11) extends the sectioning information of an AAC bitstream. This permits to detect serious errors within the spectral data of an MPEG-4 AAC bitstream. Virtual codebooks are used to limit the largest absolute value possible within a certain scalefactor band where escape values are. While referring to the same codes as codebook 11, the sixteen virtual codebooks introduced by this tool provide sixteen different limitations of the spectral values belonging to the corresponding section. Due to this, errors within spectral data resulting in spectral values exceeding the indicated limit can be located and appropriately concealed.

The Reversible Variable Length Coding tool (RVLC) replaces the Huffman and DPCM coding of the scalefactors in an AAC bitstream. The RVLC uses symmetric codewords to enable both forward and backward decoding of the scalefactor data. In order to have a starting point for backward decoding, the total number of bits of the RVLC part of the bitstream is transmitted. Because of the DPCM coding of the scalefactors, also the value of the last scalefactor is transmitted to enable backward DPCM decoding. Since not all nodes of the RVLC code tree are used as codewords, some error detection is also possible.

The Huffman codeword reordering (HCR) algorithm for AAC spectral data is based on the fact that some of the codewords can be placed at known positions so that these codewords can be decoded independent of any error within other codewords. Therefore, this algorithm avoids error propagation to those codewords, the so-called priority codewords (PCW). To achieve this, segments of known length are defined and those codewords are placed at the beginning of these segments. The remaining codewords (non-priority codewords, non-PCW) are filled into the gaps left by the PCWs using a special algorithm that minimizes error propagation to the non-PCWs codewords. This reordering algorithm does not increase the size of spectral data. Before applying the reordering algorithm itself, a pre-sorting process is applied to the codewords. It sorts all codewords depending on their importance, i. e. it determines the PCWs.

Error protection

The EP tool provides unequal error protection. It receives several classes of bits from the audio coding tools, and then applies forward error correction codes (FEC) and/or cyclic redundancy codes (CRC) for each class, according to its error sensitivity.

The error protection tool (EP tool) provides the unequal error protection (UEP) capability to the ISO/IEC 14496-3 codecs. Main features of this tool are:

- providing a set of error correcting/detecting codes with wide and small-step scalability, in performance and in redundancy
- providing a generic and bandwidth-efficient error protection framework, which covers both fixed-length frame bitstreams and variable-length frame bitstreams
- providing a UEP configuration control with low overhead

Error resilient bitstream reordering

Error resilient bitstream reordering allows the effective use of advanced channel coding techniques like unequal error protection (UEP), that can be perfectly adapted to the needs of the different coding tools. The basic idea is to rearrange the audio frame content depending on its error sensitivity in one or more instances belonging to different error sensitivity categories (ESC). This rearrangement works either data element-wise or even bit-wise. An error resilient bitstream frame is build by concatenating these instances.

Low delay

The low delay coding functionality provides the ability to extend the usage of generic low bitrate audio coding to applications requiring a very low delay of the encoding / decoding chain (e.g. full-duplex real-time communications). In contrast to traditional low delay coders based on speech coding technology, the concept of this low delay coder is based on general perceptual audio coding and is thus suitable for a wide range of audio signals. Specifically, it is derived closely from the proven architecture of MPEG-2/4 Advanced Audio Coding (AAC). Furthermore, all capabilities for coding of 2 (stereo) or more sound channels (multi-channel) are available within the low delay coder as inherited from Advanced Audio Coding.

Upstream

To allow for user on a remote side to dynamically control the streaming of the server, backchannel streams carrying user interaction information are defined.

Fine granule scalability in audio

BSAC provides fine grain scalability in steps of 1 kbit/s per audio channel, i.e. 2 kbit/s steps for a stereo signal. One base layer bitstream and many small enhancement layer bitstreams are used. In order to implement the fine grain scalability efficiently in MPEG-4 system, the fine grain audio data can be divided into the large-step layers and the large-step layers are concatenated from the several sub-frames. And the configuration of the payload transmitted over Elementary Stream (ES) can be changed dynamically depending on the environment such as the network traffic or the user interaction. So, BSAC can allow for real-time adjustments to the quality of service.

In addition to fine grain scalablity, it can improve the quality of the audio signal which is decoded from the bitstreams transmitted over error-prone channels such as mobile communication networks or Digital Audio Broadcasting (DAB)

HILN: harmonic and individual lines plus noise (parametric audio coding)

MPEG-4 parametric audio coding uses the HILN technique (Harmonic and Individual Lines plus Noise) to code non-speech signals like music at bitrates of 4 kbit/s and higher using a parametric representation of the audio signal. HILN allows independent change of speed and pitch during decoding. Furthermore HILN can be combined with MPEG-4 parametric speech coding (HVXC) to form an integrated parametric coder covering a wider range of signals and bitrates.

Silence compression for CELP

The silence compression tool comprises a Voice Activity Detector (VAD), a Discontinuous Transmission (DTX) unit and a Comfort Noise Generator (CNG) module. The tool encodes/decodes the input signal at a lower bitrates during the non-active-voice (silent) frames. During the active-voice (speech) frames, MPEG-4 CELP encoding and decoding are used.

Extension of HVXC

The operation of maximum of 4.0 kbit/s variable bitrate mode of the MPEG-4 parametric speech coder HVXC is provided in addition to the Version1 HVXC functionalities including 2.0-4.0 kbit/s fixed bitrate mode and 2.0 kbit/s maximum variable bitrate mode. Here extended operation of the variable bitrate mode with 4.0 kbit/s maximum is provided which allows higher quality variable rate coding.

Information technology — Coding of audio-visual objects —

Part 3:
Audio

Amendment 1: Audio extensions

1 Scope

ISO/IEC 14496-3 (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.

2 Normative references

The following normative documents contain provisions which, through reference in this text, constitute provisions of this Amendment. For dated references, subsequent amendments to, or revisions of, any of these publications do not apply. However, parties to agreements based on this Amendment 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.*