INTERNATIONAL STANDARD

ISO/IEC 23003-2

First edition
2010-10-01

Information technology — MPEG audio technologies —

Part 2: Spatial Audio Object Coding (SAOC)

Technologies de l'information — Technologies audio MPEG —

Partie 2: Codage d'objet audio spatial (SAOC)

Reference number
ISO/IEC 23003-2:2010(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 23003-2 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 23003 consists of the following parts, under the general title *Information technology — MPEG audio technologies*:

— *Part 1: MPEG Surround*

— *Part 2: Spatial Audio Object Coding (SAOC)*

Introduction

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.

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

The holders of these patent rights have assured 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 statements of the holders of these patent rights are registered with ISO and IEC. Information may be obtained from the companies listed in Annex G.

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 G. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Information technology — MPEG audio technologies —

Part 2:
Spatial Audio Object Coding (SAOC)

1 Scope

This part of ISO/IEC 23003 specifies the reference model of the Spatial Audio Object Coding (SAOC) technology that is capable of recreating, modifying and rendering a number of audio objects based on a smaller number of transmitted channels and additional parametric data. In the preferred modes of operating the SAOC system, the transmitted signal can be either mono or stereo. The audio objects can be represented by a mono and stereo signal or have the MPEG Surround (MPS) Multi-channel Background Object (MBO) format. The additional parametric data exhibits a significantly lower data rate than required for transmitting all objects individually, making the coding very efficient. At the same time this ensures compatibility of the transmitted signal with legacy devices.

When a multi-channel rendering setup (e.g. a 5.1 loudspeaker setup) is required, the SAOC system acts as a transcoder, converting the additional parametric data to MPS parameters, and interfaces to the MPS decoder that acts as rendering device. For certain rendering setups (e.g. a binaural or plain stereo setup), the SAOC system behaves as a decoder, using its own rendering engine. Another key feature is that the SAOC parametric data from different streams can be merged at parameter level to allow for the combination of SAOC streams, similar to the functionality of a Multi-point Control Unit (MCU).

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/IEC 13818-7:2006, *Information technology — Generic coding of moving pictures and associated audio information — Part 7: Advanced Audio Coding (AAC)*

ISO/IEC 14496-3:2009, *Information technology — Coding of audio-visual objects — Part 3: Audio*

ISO/IEC 23003-1:2007, *Information technology — MPEG audio technologies — Part 1: MPEG Surround*

3 Terms and definitions

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

3.1
audio object
input audio signal consisting of one, two or multiple channels, including Multi-channel Background Object (MBO)

3.2
frame
time segment to which SAOC processing is applied according to the data conveyed in the corresponding SAOCFrame() syntax element

3.3
hybrid filterbank
hybrid filter bank structure, consisting of a quadrature mirror filter (QMF) bank and oddly modulated Nyquist filter banks, used to transform time domain signals into hybrid subband samples

3.4
hybrid filtering
filtering step on a quadrature mirror filter (QMF) subband signal resulting in multiple hybrid subbands

NOTE The resulting hybrid subbands can be non-consecutive in frequency.

3.5
hybrid subband
subband obtained after hybrid filtering of a quadrature mirror filter (QMF) subband

NOTE The hybrid subband can have the same time/frequency resolution as a QMF subband.

3.6
input channel
input audio channel corresponding to the channels of an audio object

3.7
output channel
audio channel corresponding to a specific speaker

NOTE Channel abbreviations and loudspeaker positions are given in Table 1.

3.8
parameter band
one or more hybrid subbands applicable to one parameter

3.9
parameter time slot
specific time slot for which the parameter is defined

3.10
parameter set
parameters associated with a specific parameter time slot

3.11
parameter subset
parameters associated with a specific parameter time slot and a specific One-To-Two (OTT) box or Two-To-Three (TTT) box

3.12
processing band
one or more hybrid subbands defining the finest frequency resolution that could be controlled by the parameters

3.13
QMF bank
bank of complex exponentially modulated filters

3.14
QMF subband
subband obtained after QMF filtering of a time-domain signal, without any additional hybrid filtering stage

3.15
time segment
group of consecutive time slots

3.16
time slot
finest resolution in time for spatial audio coding (SAC) time borders

NOTE One time slot equals one subsample in the hybrid quadrature mirror filter (QMF) domain.

4 Symbols, notation and abbreviated terms

4.1 Notation

The description of the SAOC system uses the following notation:

- Vectors are indicated by bold lower-case names, e.g. **vector**.
- Matrices (and vectors of vectors) are indicated by bold upper-case single letter names, e.g. **M**.
- Variables are indicated by italic, e.g. *variable*.
- Functions are indicated as *func(x)*.

For equations (and flowcharts), normal mathematical (and pseudo-code) interpretation is assumed with no rounding or truncation unless explicitly stated.

4.2 Operations

4.2.1 Scalar operations

X^* is the complex conjugate of X.

$y = \log_{10}(x)$ is the base-10 logarithm of x.

$y = \min(\ldots,\ldots)$ is the minimum value in the argument list.

$y = \max(\ldots,\ldots)$ is the maximum value in the argument list.

4.2.2 Vector and matrix operations

$\mathbf{m} = diag(\mathbf{M})$ is diagonal of matrix $\mathbf{M}$.

$\mathbf{y} = sort(\mathbf{x})$ is equal to the sorted vector $\mathbf{x}$, where the elements of $\mathbf{x}$ are sorted in ascending order.

$y = trace(\mathbf{M})$ is sum of all diagonal elements of matrix $\mathbf{M}$.

4.3 Constants

ε is an additive constant to avoid division by zero, e.g. $\varepsilon = 10^{-9}$.

4.4 Variables

$a_{i,y}^{l,m}$ is the virtual speaker transfer function, defined for binaural output channel i, audio object y and all parameter time slots l and processing bands m.

$\mathbf{D}$ is the downmix matrix.

$\mathbf{D}_{\text{CLD}}$ is the three dimensional matrix holding the dequantized, and mapped CLD data for every OTT box, every parameter set, and M_{proc} bands.

$\mathbf{D}_{\mathrm{ICC}}$ is the three dimensional matrix holding the dequantized, and mapped ICC data for every OTT or TTT box, every parameter set, and M_{proc} bands.

$\mathbf{D}_{\mathrm{CPC\_1}}$, $\mathbf{D}_{\mathrm{CPC\_2}}$ are the three dimensional matrices holding the dequantized, and mapped first and second CPC data for every TTT box, every parameter set, and M_{proc} bands.

$\mathbf{D}_{\mathrm{CLD\_1}}$, $\mathbf{D}_{\mathrm{CLD\_2}}$ are the three dimensional matrices holding the dequantized, and mapped first and second CLD data for every TTT box, every parameter set, and M_{proc} bands.

$\mathbf{D}_{\mathrm{DCLD}}$ is the two dimensional matrix holding the dequantized, and mapped DCLD data for every input channel, and every parameter set.

$\mathbf{D}_{\mathrm{DMG}}$ is the two dimensional matrix holding the dequantized, and mapped DMG data for every input channel, and every parameter set.

$\mathbf{D}_{\mathrm{IOC}}$ is the four dimensional matrix holding the dequantized, and mapped IOC data for every input channel pair, every parameter set, and M_{proc} bands.

$\mathbf{D}_{\mathrm{NRG}}$ is the two dimensional matrix holding the dequantized, and mapped NRG data for the highest energy within every parameter set, and M_{proc} bands.

$\mathbf{D}_{\mathrm{OLD}}$ is the three dimensional matrix holding the dequantized, and mapped OLD data for every input channel, every parameter set, and M_{proc} bands.

$\mathbf{D}_{\mathrm{PDG}}$ is the three dimensional matrix holding the dequantized, and mapped PDG data for every downmix channel, every parameter set, and M_{proc} bands.

$H^{m}_{i,\{L,R\}}$ is the HRTF parameter which represents the average level with respect to the left and right ear {*L*, *R*} for the HRTF database index i and all processing bands m.

$\mathbf{idx}XXX(\ldots,\ldots)$ is a three dimensional matrix holding the Huffman and delta decoded indices. *XXX* can be any of OLD, IOC, NRG, DCLD, DMG, PDG.

K is the number of hybrid subbands.

L is the number of parameter sets.

M is the number of downmix channels.

M_{proc} is the number of processing bands.

M_{QMF} is the number of QMF subbands depending on sampling frequency.

$\mathbf{M}^{l,m}$ is the OTN/TTN upmix matrix for the prediction mode of operation

$\mathbf{M}^{l,m}_{Energy}$ is the OTN/TTN upmix matrix for the energy mode of operation

$\mathbf{M}^{n,k}_{1}$, $\mathbf{M}^{n,k}_{2}$ are the time and frequency variant pre-matrices, defined for all time slots n and all hybrid subbands k.

$\mathbf{M}^{l,m}_{\mathrm{ren}}$ is the time and frequency variant rendering matrix, defined for all parameter time slots l and all processing bands m.

N is the number of SAOC input channels of audio objects.

N_{EAO} is the number of EAO channels.

N_{MPS} is the number of MPS output channels.

N_{HRTF} is the number of different HRTFs in the HRTF database.

P frame length.

$\mathbf{W}_{\mathrm{ADG}}^{l,m}$ is the time and frequency variant matrix including ADGs, defined for all parameter time slots l and all processing bands m.

$\mathbf{W}_{h}^{l,m}$ is the time and frequency variant sub-rendering matrix, defined for OTT box h (of the MPS “5-1-5” tree-structure), all parameter time slots l and all processing bands m.

$\mathbf{W}_{\mathrm{PDG}}^{l,m}$ is the time and frequency variant matrix including PDGs, defined for all parameter time slots l and all processing bands m.

$\mathbf{s}^{n,k}$ is a vector with the hybrid subband (encoder) input channels, defined for all time slots n and all hybrid subbands k.

$\mathbf{x}^{n,k}$ is a vector with the hybrid subband (transcoder/decoder) input signals (downmix and residuals), defined for all time slots n and all hybrid subbands k.

$\mathbf{y}^{n,k}$ is a vector with the (transcoder/decoder) output hybrid subband signals, which are fed into the hybrid synthesis filter banks, defined for all time slots n and all hybrid subbands k.

ϕ_i^m is the HRTFs parametric representation of the average phase difference, defined for the HRTF database index i and all processing bands m.

4.5 Abbreviated terms

ADG Arbitrary Downmix Gain

CLD Channel Level Difference, describes the energy difference between two channels

CPC Channel Prediction Coefficient, used for recreating three or more channels from two channels

DCLD Downmix Channel Level Difference describes the gain differences of objects contributing to the left and right downmix channel in case of a stereo downmix

DCU Distortion Control Unit

DMG DownMix Gain, gains applied to each object before downmixing

EAO Enhanced Audio Object

HRTF Head Related Transfer Function

ICC Inter Channel Correlation, describes the correlation between two channels

IOC Inter Object Correlation, describes the correlation between two channels of audio objects

LD Low Delay

MBO Multi-channel Background Object

MCU Multi-point Control Unit

MPS MPEG Surround

N/A Not Applicable

NRG absolute object eNeRGy, specifies the absolute energy of the object with the highest energy for the corresponding frequency band

OLD Object Level Difference, describes intensity differences between one object and the object with the highest energy for the corresponding frequency band

OTN conceptual “One-To-N” unit that takes one channel as input and produces N channels as output

OTT conceptual “One-To-Two” unit that takes one channel as input and produces two channels as output

PDG Post(processing) Downmix Gains, describes intensity differences between the encoder-generated downmix and the post(processed) downmix for the corresponding frequency band

QMF Quadrature Mirror Filter

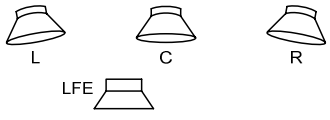
SAC Spatial Audio Coding

SAOC Spatial Audio Object Coding

TTN conceptual "Two-To-N" unit that takes two channels as input and produces N channels as output

TTT conceptual "Two-To-Three" unit that takes two channels as input and produces three channels as output

Table 1 – Channel abbreviations and loudspeaker positions

Channel abbreviation	Loudspeaker position	Figure
L	Left Front	
R	Right Front	
C	Center Front	
LFE	Low Frequency Enhancement	
Ls	Left Surround	
Rs	Right Surround	

5 SAOC overview

5.1 Introduction

Spatial Audio Object Coding (SAOC) is a parametric multiple object coding technique. It is designed to transmit a number of audio objects in an audio signal that comprises M channels. Together with this backwards compatible downmix signal, object parameters are transmitted that allow for recreation and manipulation of the original object signals. An SAOC encoder produces a downmix of the object signals at its input and extracts these object parameters. The number of objects that can be handled is in priniciple not limited.

The object parameters are quantized and coded efficiently into an SAOC bitstream.

The downmix signal can be compressed and transmitted without the need to update existing coders and infrastructures. The object parameters, or SAOC side information, are transmitted in a low bitrate side channel, e.g. the ancillary data portion of the downmix bitstream.

On the decoder side, the input objects are reconstructed and at the same time rendered to a certain number of playback channels. The rendering information containing reproduction level and panning position for each object is user supplied or can be extracted from the SAOC bitsream (e.g. preset information). The rendering information can be time variant. Output scenarios can range from mono to multi-channel (e.g. 5.1) and are independent from both, the number of input objects and the number of downmix channels. Binaural rendering of objects is possible including azimuth and elevation of virtual object positions. An optional effects interface allows for advanced manipulation of object signals, besides level and panning modification.

The objects themselves can be mono signals, stereophonic signals, as well as multi-channel signals (e.g. 5.1 channels). Typical downmix configurations are mono and stereo.

5.2 Basic structure of the SAOC transcoder/decoder

The SAOC transcoder/decoder module described below may act either as a stand-alone decoder or as a transcoder from an SAOC to an MPS bitstream, depending on the intended output channel configuration. The following table illustrates the differences between the two modes of operation:

Table 2 – Operation modes of the SAOC

Output signal configuration	# of output channels	SAOC module mode	SAOC module output	MPS decoder required
mono/stereo/binaural	1 or 2	Decoder	PCM output	No
multi-channel configuration	> 2	Transcoder	MPS bitstream, downmix signal	Yes

Figure 1 shows the basic structure of the SAOC transcoder/decoder architecture. The residual processor extracts the EAOs from the incoming downmix using the residual information contained in the SAOC bitstream. The downmix pre-processor processes the regular audio objects. The EAOs and processed regular audio objects are combined to the output signal for the SAOC decoder mode or to the MPS downmix signal for the SAOC transcoder mode. The detailed descriptions of these processing blocks are given in the corresponding subclauses, namely, 7.6 and 7.7 describe the SAOC transcoder/decoder functionality and 7.8 explains handling of extended audio objects and residual processing.

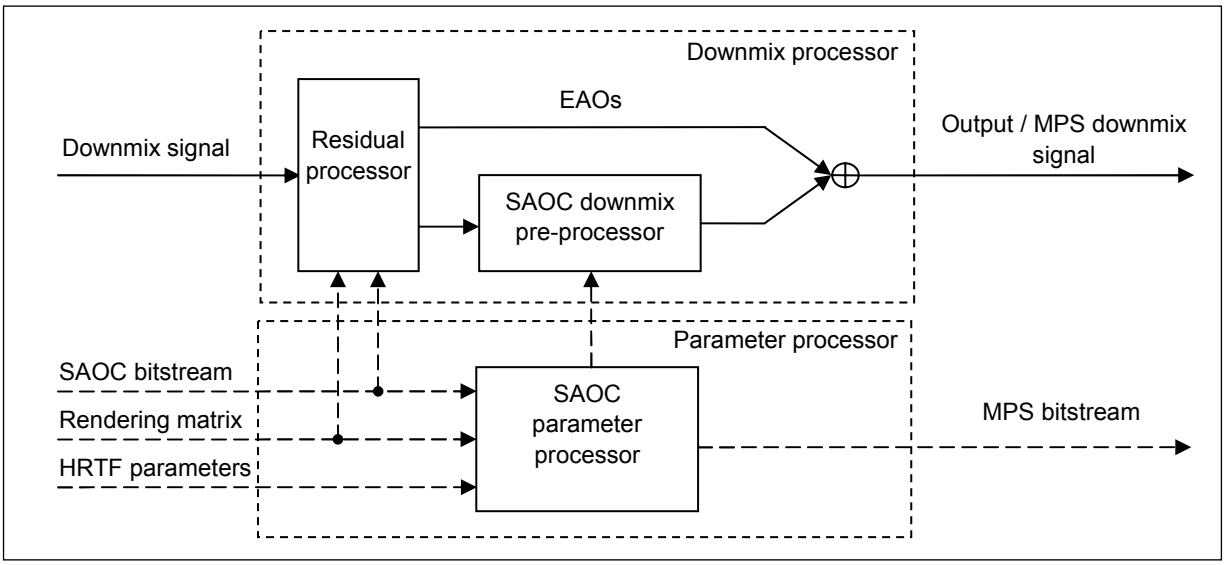


Figure 1 – Overall structure of the SAOC transcoder/decoder architecture

Figure 2 (left) shows a block diagram of an SAOC transcoder unit. It consists of an SAOC parameter processor and a downmix processor module. The SAOC parameter processor decodes the SAOC bitstream and has furthermore a user interface from which it receives additional input in form of generally time variant rendering information. It provides steering information for the downmix processor. The SAOC transcoder outputs an MPS bitstream and downmix signal, as an input to the MPS decoder. In case of a mono downmix, the downmix pre-processor leaves the downmix signal unchanged. However, in case of a stereo downmix, it is functional to pre-process the downmix signal to allow more flexible object panning than is supported by the MPS rendering engine alone. In case of a mono/stereo/binaural output configuration the SAOC system works in decoder mode and MPS decoding is omitted, see Figure 2 (right). Here the downmix processing module directly provides the output signal.

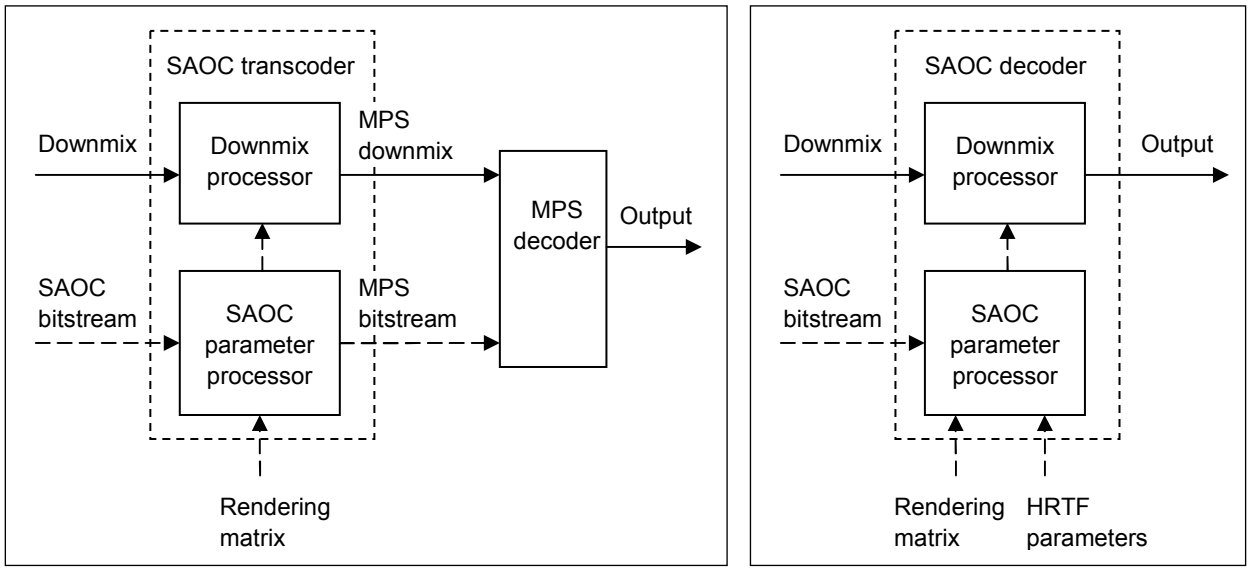


Figure 2 – Block diagrams of the SAOC transcoder (left) and decoder (right) processing modes

5.3 Tools and functionality

5.3.1 General SAOC tools

5.3.1.1 Introduction

The SAOC system incorporates a number of tools that allow for flexible complexity and/or quality trade-off, as well as a diverse set of functionality. In the following subclauses some key-features of SAOC are briefly outlined.

5.3.1.2 Binaural decoding

The SAOC system can be operated in a binaural mode. This enables a multi-channel impression over headphones by means of Head Related Transfer Function (HRTF) filtering.

5.3.1.3 Efficient multipoint control unit support

In order to use the SAOC concept for teleconferencing applications a Multipoint Control Unit (MCU) functionality of combining the signals of several communication partners without decoding/re-encoding the corresponding audio objects is provided. The MCU combines the input SAOC side information streams into one common SAOC bitstream in a way that the parameters representing all audio objects from the input bitstreams are included in the resulting output bitstream. These calculations are performed in the parameter domain without the need to analyze the downmix signals and, therefore, introduce no additional delay in the signal processing chain.

5.3.1.4 External downmix

The SAOC system is capable of handling not only encoder-generated downmixes but also post(processed) downmixes supplied to the encoder in addition to the input audio object signals. In this case, Post Downmix Gains (PDGs) are calculated in the encoder and conveyed as a part of the SAOC bitstream. The difference of the downmix signals is compensated for at the SAOC decoder side.

5.3.1.5 Multichannel background object

The audio input to a SAOC encoder can contain a so-called Multi-channel Background Object (MBO). Generally, the MBO can be considered as a complex sound scene involving a large and often unknown number of sound sources, for which no controllable rendering functionality is required. The MBO is represented by a downmix of the MPS encoded complex sound scene and corresponding MPS parameters.

5.3.1.6 Enhanced audio object processing

A special "Karaoke-type" application scenario requires a total suppression of specific objects, typically the lead vocals, while keeping the perceptual quality of the background sound scene unharmed. High sound quality is assured by the incorporation of residual coding enabling a better separation of the background object and foreground objects. The current EAO processing mode supports reproduction of both EAO and regular objects exclusively and arbitrary mixtures of these object groups.

5.3.1.7 Distortion control unit

The distortion control unit is incorporated into the SAOC system in order to provide a flexible control for users and audio content providers over the SAOC rendering functionality and audio output quality.

5.3.1.8 Predefined rendering information

The SAOC system is capable of starting playback with some initial predefined settings which can be stored and/or transmitted in SAOC bitstream. These settings can be dynamically updated. The SAOC system allows instantaneous switching between them if more than one set of predefined settings is available.