INTERNATIONAL STANDARD

ISO/IEC 23003-2

Second edition
2018-12

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 23003-2:2018(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: 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.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see http://patents.iec.ch).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see: www.iso.org/iso/foreword.html.

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

This second edition cancels and replaces the first edition (ISO/IEC 23003-2:2010), which has been technically revised. It also incorporates the Amendments ISO/IEC 23003-2:2010/Amd 1:2015, ISO/IEC 23003-2:2010/Amd 2:2015, ISO/IEC 23003-2:2010/Amd 3:2015, ISO/IEC 23003-2:2010/Amd 4:2016 and ISO/IEC 23003-2:2010/Amd 5:2016 and the Technical Corrigenda ISO/IEC 23003-2:2010/Cor 1:2012 and ISO/IEC 23003-2:2010/Cor 2:2014.

The main changes compared to the previous edition are as follows:

— clarifications on SAOC-DE profile description;

— corrections to SAOC-DE profile specification;

— corrections to SAOC-DE profile;

— corrections to MPEG SAOC IS text;

— corrections to the low power mode.

A list of all parts in the ISO/IEC 23003 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

In the preferred modes of operating, the SAOC system, the transmitted signal can be either mono, stereo or 3-channel. The audio objects can be represented by a mono, stereo, or 3-channel 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).

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 a patent.

ISO and IEC take no position concerning the evidence, validity and scope of this patent right. The holder of this patent right has assured ISO and IEC that he/she is 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 ISO and IEC. Information may be obtained from:

Qualcomm Incorporated
6455 Lusk Blvd
US-San Diego, CA 92121-2779

Fraunhofer Institute for Integrated Circuits IIS
Leonrodstrasse 68
DE-80636 München

LG Electronics
16 Woomyeon-Dong Seocho-Gu
KR-Seoul 137-724

Koninklijke Philips Electronics N.V.
High Tech Campus 44
NL-5656 AE , Eindhoven

Electronics and Telecommunications Research Institute
161 Gajeong-dong Yuseong-gu
KR-Daejeon 305-350

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 above. 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 document 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.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements 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 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.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— IEC Electropedia: available at http://www.electropedia.org/

— ISO Online browsing platform: available at http://www.iso.org/obp

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 (3.15) to which SAOC processing is applied according to the data conveyed in the corresponding SAOCFrame() or SAOCDEFrame() syntax elements

3.3
hybrid filterbank
structure, consisting of a quadrature mirror filter (QMF) bank and oddly modulated Nyquist filter banks, used to transform time domain signals into *hybrid subband* (3.5) samples

3.4
hybrid filtering
filtering step on a quadrature mirror filter (QMF) subband signal resulting in multiple *hybrid subbands* (3.5)

Note 1 to entry: The resulting hybrid subbands can be non-consecutive in frequency.

3.5
hybrid subband
subband obtained after *hybrid filtering* (3.4) of a quadrature mirror filter (QMF) subband

Note 1 to entry: 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* (3.1)

3.7
output channel
audio channel corresponding to a specific speaker

Note 1 to entry: Channel abbreviations and loudspeaker positions are given in Table 1.

3.8
parameter band
one or more *hybrid subbands* (3.5) applicable to one parameter

3.9
parameter time slot
specific *time slot* (3.16) for which the parameter is defined

3.10
parameter set
parameters associated with a specific *parameter time slot* (3.9)

3.11
parameter subset
parameters associated with a specific *parameter time slot* (3.9) and a specific one-to-two (OTT) box or two-to-three (TTT) box

3.12
processing band
one or more *hybrid subbands* (3.5) 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)

3.16
time slot
finest resolution in time for spatial *audio object* (3.1) coding (SAOC) time borders

Note 1 to entry: One time slot equals one subsample in the hybrid quadrature mirror filter (QMF) domain.

4 Notations and abbreviated terms

4.1 Notation

The description of the SAOC system uses the following notations:

— 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.
$y = \min(...,...)$	is the minimum value in the argument list.
$y = \max(...,...)$	is the maximum value in the argument list.
$\exp(x)$	is the exponential function of x.

4.2.2 Vector and matrix operations

$\mathbf{m} = diag(\mathbf{M})$	is main diagonal of matrix, **M**.
$\mathbf{y} = sort(\mathbf{x})$	is equal to the sorted vector **x**, where the elements of **x** are sorted in ascending order.
$y = trace(\mathbf{M})$	is sum of all diagonal elements of matrix, **M**.
$\mathbf{M}^*$	is the complex conjugate transpose of **M**.

4.3 Constants

ε	is a constant to avoid division by and logarithm of zero, e.g. $\varepsilon = 10^{-9}$.
$\mathbf{0}_{A \times B}$	is a matrix of size $A \times B$ consisting of zeros.
$\mathbf{I}_A$	is an identity matrix of size $A \times A$.

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.

$\mathbf{D}_{CLD}$	is the three-dimensional matrix holding the dequantized, and mapped CLD data for every OTT box, every parameter set and M_{proc} bands.
$\mathbf{D}_{ICC}$	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.
$\mathbf{D}_{CPC\_1}$, $\mathbf{D}_{CPC\_2}$	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.
$\mathbf{D}_{CLD\_1}$, $\mathbf{D}_{CLD\_2}$	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.
$\mathbf{D}_{DCLD}$	is the matrix holding the dequantized, and mapped DCLD data for every input channel and every parameter set.
$\mathbf{D}_{DMG}$	is the matrix holding the dequantized, and mapped DMG data for every input channel and every parameter set. If DMG data contains information for more than one downmix channel, $\mathbf{D}_{DMG}$ is a three-dimensional matrix holding the dequantized, and mapped DMG data for every input channel, every downmix channel and every parameter set.
$\mathbf{D}_{IOC}$	is the four-dimensional matrix holding the dequantized, and mapped IOC data for every input channel pair, every parameter set and M_{proc} bands.
$\mathbf{D}_{NRG}$	is the two-dimensional matrix holding the dequantized, and mapped NRG data for the highest energy within every parameter set and M_{proc} bands.
$\mathbf{D}_{OLD}$	is the three-dimensional matrix holding the dequantized, and mapped OLD data for every input channel, every parameter set and M_{proc} bands.
$\mathbf{D}_{PDG}$	is the three-dimensional matrix holding the dequantized, and mapped PDG data for every downmix channel, every parameter set and M_{proc} bands.
$\mathbf{D}_{BGO}$	is the downmix sub-matrix for BGOs.
$\mathbf{D}_{FGO}$	is the downmix sub-matrix for FGOs.
$H^{m}_{i,\{L,R\}}$	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*.
idx*XXX*(...,...)	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.

$\mathbf{M}^{l,m}_{\text{Energy}}$ 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*.

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

$\mathbf{G}^{l,m}_{\text{DE}}$ is the time and frequency variant parametric processing matrix, defined for all parameter time slots, *l*, and all processing bands, *m*.

$\mathbf{M}^{l,m}_{\text{DE}}$ is the time and frequency variant residual processing matrix, defined for all parameter time slots, *l*, and all processing bands, *m*.

m_{BGO} is the modification gain for BGOs.

m_{FGO} is the modification gain for FGOs.

m_{G} is the decoder limited modification gain.

m^{input}_{G} is the input modification gain.

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

N_{FGO} is the number of FGOs.

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.

N_{g} is the number of groups of downmix signals.

N^q_{g} is the number of downmix signals assigned to group $\mathbf{g}_q$, defined for all group indices, *q*.

$\mathbf{g}_q$ is a vector with the indices of the downmix signals assigned to the same group, defined for all group indices, *q*.

P is the frame length.

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

$\mathbf{W}^{l,m}_h$ 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

BGO background object

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

DE dialogue enhancement

DMG downmix gain; gains applied to each object before downmixing

EAO enhanced audio object

FGO foreground 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 parameter band

OLD object level difference, describes intensity differences between one object and the object with the highest energy for the corresponding parameter 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 parameter band

QMF quadrature mirror filter

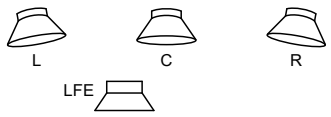
SAC spatial audio coding

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 General

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 principle 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. Table 2 illustrates the differences between the two modes of operation.

Table 2 — Operation modes of the SAOC

Output signal configuration	# of output channels	# of input channels	SAOC module mode	SAOC module output	MPS decoder required
Mono/stereo/binaural/ 3-channel configuration	1, 2 or 3	1, 2 or 3	Decoder	PCM output	No
Multi-channel configuration	>2	1 or 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, 7.5 and 7.5.4, which describe the SAOC transcoder/decoder functionality and 7.6 explains handling of enhanced audio objects and residual processing.

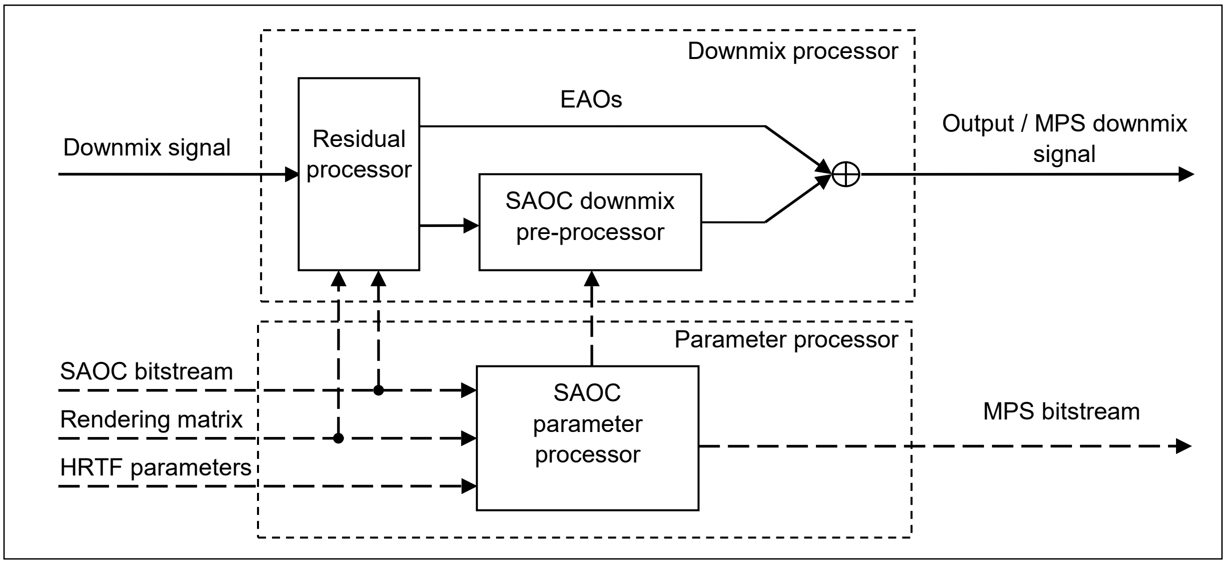


Figure 1 — Overall structure of the SAOC transcoder/decoder architecture

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/multi-channel 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.

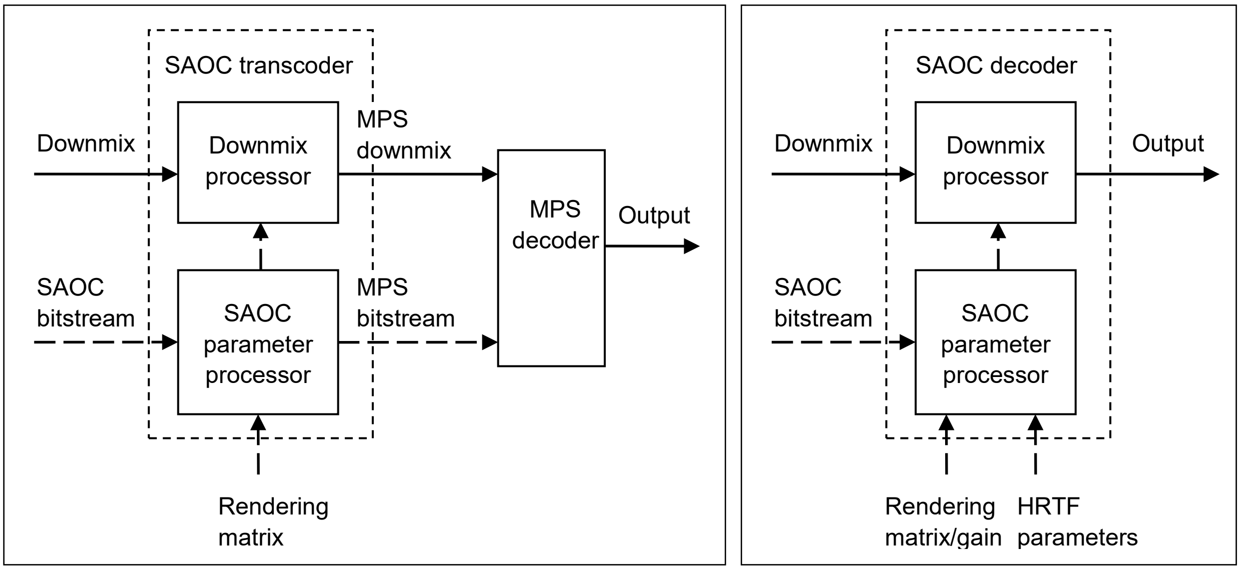


Figure 2 — Block diagrams of the SAOC transcoder (left) and decoder (right) processing modes