INTERNATIONAL STANDARD

ISO/IEC 23003-1

First edition
2007-02-15

Information technology — MPEG audio technologies —

Part 1: MPEG Surround

Technologies de l'information — Technologies audio MPEG —

Partie 1: Ambiance MPEG

Reference number
ISO/IEC 23003-1:2007(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. 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 23003-1 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*

Introduction

The following MPEG International Standard describes the coding of multi-channel signals based on a downmixed signal of the original multi-channel signal, and associated spatial parameters. It offers lowest possible data rate for coding of multi-channel signals, as well as an inherent mono or stereo downmix signal included in the data stream. Hence, a mono or stereo signal can be expanded to multi-channel by a very small additional data overhead. Furthermore, the International Standard describes binaural decoding of the MPEG Surround stream, enabling a surround sound experience over headphones. The International Standard furthermore defines an Enhanced Matrix Mode that enables a multi-channel upmix from a stereo signal without any spatial parameters.

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.

The 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 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 the ISO and IEC. Information may be obtained from the companies listed in Annex J.

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

Information technology — MPEG audio technologies —

Part 1:
MPEG Surround

1 Scope

This International Standard describes the MPEG Surround standard (Spatial Audio Coding, SAC), that is capable of re-creating N channels based on M<N transmitted channels, and additional control data. In the preferred modes of operating the spatial audio coding system, the M channels can either be a single mono channel or a stereo channel pair. The control data represents a significant lower data rate than required for transmitting all N channels, making the coding very efficient while at the same time ensuring compatibility with both M channel devices and N channel devices.

This International Standard incorporates a number of tools enabling a number of features that allow for broad application of the International Standard. A key feature is the ability to scale the spatial image quality gradually from very low spatial overhead towards transparency. Another key feature is that the compatible decoder input can be made compatible to existing matrix surround technologies. All tools are grouped to cover certain profiles.

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

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

3 Terms and definitions

3.1 Definitions

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

3.1.1
channel
input or output audio channel corresponding to a specific speaker, as given by Table 1 — and illustrated in Figure 1

Table 1 — Channel abbreviation and loudspeaker position

Channel abbreviation	Loudspeaker position
L	Left Front
R	Right Front
C	Center Front
LFE	Low Frequency Enhancement
Ls	Left Surround
Rs	Right Surround
Lc	Left Front Center
Rc	Right Front Center
Lsr	Rear Surround Left
Rsr	Rear Surround Right
Cs	Rear Center
Lsd	Left Surround Direct
Rsd	Right Surround Direct
Lss	Left Side Surround
Rss	Right Side Surround
Lw	Left Wide Front
Rw	Right Wide front
Lv	Left Front Vertical Height
Rv	Right Front Vertical Height
Cv	Center Front Vertical Height
Lvr	Left Surround Vertical Height Rear
Rvr	Right Surround Vertical Height Rear
Cvr	Center Vertical Height Rear
Lvss	Left Vertical Height Side Surround
Rvss	Right Vertical Height Side Surround
Ts	Top Center Surround
LFE2	Low Frequency Enhancement 2

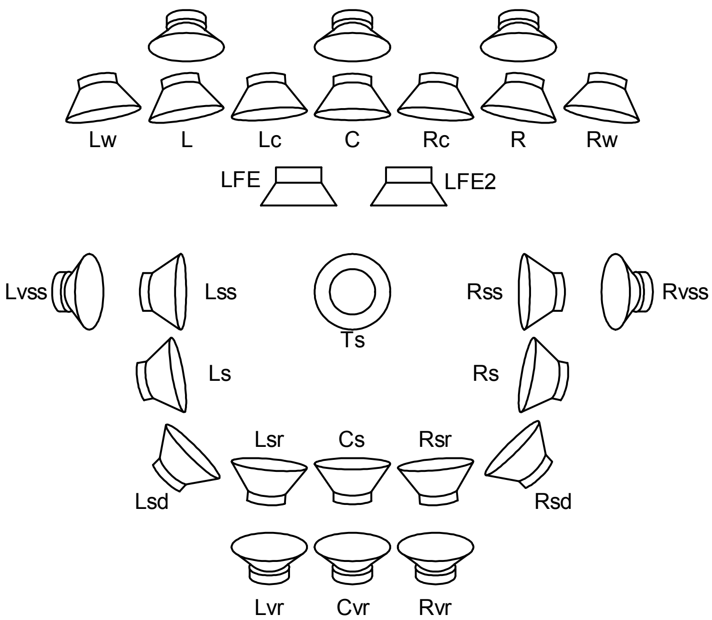

Figure 1 — Loudspeaker positions

3.1.2
Channel Level Difference
CLD
energy difference between two channels

3.1.3
CLD band
one or more hybrid subbands for which a single CLD parameter applies

3.1.4
Channel Prediction Coefficient
CPC
prediction coefficient used for re-creating three channels from two channels

3.1.5
CPC band
one or more hybrid subbands for which a single CPC parameter applies

3.1.6
hybrid filterbank
hybrid filter bank structure, consisting of a QMF bank and oddly modulated Nyquist filter banks, used to transform time domain signals into hybrid subband samples

3.1.7
hybrid filtering
filtering step on a QMF subband signal resulting in multiple hybrid subbands

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

3.1.8
hybrid subband
subband obtained after hybrid filtering of a QMF subband

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

3.1.9
Inter Channel Correlation
ICC
correlation or coherence between two channels

3.1.10
ICC band
one or more hybrid subbands for which a single IIC parameter applies

3.1.11
NA
Not Applicable

3.1.12
M-N-M configuration
configuration of the spatial audio coding system that re-creates M channels from N downmixed channel and the corresponding spatial parameters, e.g. 5-1-5 configuration or 5-2-5 configuration

3.1.13
OTT box
conceptual one-to-two box that takes one channel as input and produces two channels as output

3.1.14
parameter band
one or more hybrid subbands applicable to one parameter

3.1.15
parameter band border
parameter band delimiter, expressed as a specific hybrid subband

3.1.16
parameter time slot
specific time slot for which the parameter is defined

3.1.17
parameter set
parameters associated with a specific parameter time slot

3.1.18
parameter subset
parameters associated with a specific parameter time slot and a specific OTT box or TTT box

3.1.19
processing band
one or more hybrid subbands defining the finest frequency resolution that could be controlled by the parameters

3.1.20
QMF bank
bank of complex exponentially modulated filters

3.1.21
QMF subband
subband obtained after QMF filtering of a time-domain signal, without any additional hybrid filtering stage

3.1.22
SAC
Spatial Audio Coder

3.1.23
SAC frame
time segment to which processing is applied according to the data conveyed in the corresponding SpatialFrame() syntax element

3.1.24
time segment
group of consecutive time slots

3.1.25
time slot
finest resolution in time for SAC time borders

NOTE One time slot equals one subsample in the hybrid QMF domain.

3.1.26
TTT box
conceptual two-to-three box that takes two channels as input and produces three channels as output

3.2 Notation

The description of the Spatial Audio Coder 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)*.
- Real numbers are denoted by R
- Complex numbers are denoted by C

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

3.3 Operations

3.3.1 Scalar operations

X^* is the complex conjugate of X.

$y = \lfloor x \rfloor$ represents rounding down to the nearest integer, i.e., the largest integer number that is less than x.

or $y = \lceil x \rceil$ represents rounding up to the nearest integer, i.e., the smallest integer number that is not less than x.

$y = INT(x)$ represents truncation to integer (only keep the integer part), i.e., conversion to the integer number with the same sign as x and with an absolute value smaller than or equal to the absolute value of x.

$y = \log_2(x)$ is the base-2 logarithm of x.

$y = \log_{10}(x)$ is the base-2 logarithm of x.

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

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

$y = \mathrm{mod}(x,z)$ is the modulo operation y = (x - n*z) where n = ceil(x/z)-1 defined for z ~= 0..

$y = \mathrm{round}(x)$ represents rounding to the nearest integer. Halfway cases are rounded away from zero.

$y = \mathrm{sign}(x)$ the sign of x, hence defined as -1 for negative values of x, 1 for positive values and 0 for 0.

3.3.2 Vector operations

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

$y = length(\mathbf{x})$. y is the number of elements of the vector $\mathbf{x}$.

3.4 Constants

ε	A constant to avoid division by zero, e.g. 96 dB below maximum signal input.

3.5 Variables

$\mathbf{a}^{m}(l)$	aliasing condition vector defined for every parameter time slot l and all QMF subbands m that are the last subband (highest in frequency) within a parameter band.
ch	is the current audio channel.
$\mathbf{D}_{\mathrm{ATD}}$	is the three dimensional matrix holding arbitrary tree data, i.e. mapped CLD data, for every OTT box, every parameter set, and M_{proc} bands, for the arbitrary tree.
$\mathbf{D}_{\mathrm{CLD}}$	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.
$\mathbf{D}_{\mathrm{CPC\_1}}$	is the three dimensional matrix holding the dequantized, and mapped first CPC data for every TTT box, every parameter set, and M_{proc} bands.
$\mathbf{D}_{\mathrm{CPC\_2}}$	is the three dimensional matrix holding the dequantized, and mapped second CPC data for every TTT box, every parameter set, and M_{proc} bands.
$\mathbf{D}_{\mathrm{CLD\_1}}$	is the three dimensional matrix holding the dequantized, and mapped first CLD data for every TTT box, every parameter set, and M_{proc} bands.
$\mathbf{D}_{\mathrm{CLD\_2}}$	is the three dimensional matrix holding the dequantized, and mapped second CLD data for every TTT box, every parameter set, and M_{proc} bands.
$\mathbf{D}_{YYY}^{\mathrm{Q}}$	is a three dimensional matrix similar to $\mathbf{D}_{YYY}$, the data is dequantized as for $\mathbf{D}_{YYY}$, however it has M_{par} bands of data. YYY can be any of CLD, ICC, CPC_1, CPC_2, CLD_1, CLD_2 or ATD.
$\mathbf{envRatio}_{X}$	is a vector with GES envelope data for each channel X.
Fs	is the sampling frequency of the Spatial Audio Tool.
$\mathbf{G}$	is a three dimensional matrix holding the dequantized and mapped gain correction data for all input channels, parameter set, and M_{proc} bands.
$\mathbf{G}^{\mathrm{Q}}$	is a three dimensional matrix holding the dequantized gain correction data for all input channels, parameter set, and M_{par} bands.
$\mathbf{idx}XXX(,,)$	is a three dimensional matrix holding the Huffman and delta decoded indices. XXX can be any of CLD, ICC, CPC_1, CPC_2, CLD_1, CLD_2 or ATD.
K	number of hybrid subbands, 71.
K_c	number of complex QMF subbands for Low Power MPEG Surround, $K_c = 8$.
L	number of parameter sets.

M_{proc}	is the number of processing bands, 28.
M^{c}_{proc}	number of complex processing bands for Low Power MPEG Surround, 12.
M_{par}	is the number of parameter bands signalled by *bsFreqRes*.
M_{QMF}	is the number of QMF subbands depending on sampling frequency as defined in subclause 6.3.3.
$\mathbf{m}_{\text{resPar}}$	is a vector with the number of parameter bands that each residual cover.
$\mathbf{m}_{\text{resProc}}$	is a vector with the number of processing bands that each residual cover.
$\mathbf{m}_{\text{tttLowProc}}$	is a vector with the number of processing bands for the low range in the TTT boxes.
$\mathbf{m}_{\text{tttHighProc}}$	is a vector with the number of processing bands for the high range in the TTT boxes.
$\mathbf{M}_1^{n,m}$	is the time and frequency variant pre- matrix, defined for all time slots n and all hybrid subbands m.
$\mathbf{M}_2^{n,m}$	is the time and frequency variant mix-matrix, defined for all time slots n and all hybrid subbands m.
$\mathbf{r}^{m}(l)$	weighted correlation sum based on the input downmix signal, defined for every parameter time slot l and all QMF subbands m that have an adjoining parameter border, used for Low Power MPEG surround.
reset	a variable (in the encoder and the decoder) set to one if certain data elements have changed from the previous frame, otherwise set to zero.
$\mathbf{S}_{\text{proc}}$	a matrix indicating for every parameter set and processing band if smoothing is applied.
$\mathbf{s}_{\text{delta}}$	a vector indicating for every time-slot the smoothing filter coefficient.
t	is of length *L* and contains parameter time slots for all CLD, ICC, and CPC parameter sets in the current frame.
$\mathbf{Tree}(ch,,)$	a 3 dimensional matrix, which for each input channel to the Arbitrary Tree have a column for each output signal of the sub-tree indexing the OTT modules the input signal must pass before the output is reached.
$\mathbf{Tree}_{\text{sign}}(ch,,)$	a 3 dimensional matrix, which for each input channel to the Arbitrary Tree have a column for each output signal of the sub-tree indicating whether the upper (1) or the lower (-1) output of an OTT module should be followed to reach the output signal.
$\mathbf{Tree}_{\text{depth}}(ch,)$	a matrix which for each input channel to the Arbitrary Tree have the number of OTT modules that are passed for every output channel.
$\mathbf{Tree}_{\text{outChan}}(ch)$	is a vector with *numOutChan* elements and each element contain the number of output channels for each Arbitrary Sub-tree.

$\mathbf{v}^{n,m}$	is a vector with the hybrid subband output from the pre gain matrix $\mathbf{M}_1^{n,m}\mathbf{x}^{n,m}$, defined for all time slots n and all hybrid subbands m.
$\mathbf{w}^{n,m}$	is a vector with the hybrid subband output from the decorrelators, the pre-gain matrix and residuals , defined for all time slots n and all hybrid subbands m.
$\mathbf{w}_{\text{diffuse}}^{n,m}$	is a vector with the hybrid subband output from the decorrelators, defined when temporal shaping is used, defined for all time slots n and all hybrid subbands m.
$\mathbf{w}_{\text{direct}}^{n,m}$	is a vector with the hybrid subband output from the decorrelators, the pre-gain matrix and residuals, defined when temporal shaping is used, defined for all time slots n and all hybrid subbands m.
$\mathbf{x}^{n,m}$	is a vector with the hybrid subband input signals (down-mix and residuals), defined for all time slots n and all hybrid subbands m.
$\mathbf{y}^{n,m}$	is a vector with the output hybrid subband signals, which are feed into the hybrid synthesis filter banks, defined for all time slots n and all hybrid subbands m.
$\mathbf{y}_{\text{diffuse}}^{n,m}$	is a vector with the output hybrid subband signals for the diffuse part of the output signal, which is defined when temporal processing is applied, defined for all time slots n and all hybrid subbands m.
$\mathbf{y}_{\text{direct}}^{n,m}$	is a vector with the output hybrid subband signals for the direct part of the output signal, which is defined when temporal processing is applied, defined for all time slots n and all hybrid subbands m.

4 MPEG Surround overview

4.1 Introduction

MPEG Surround is based on a principle called Spatial Audio Coding (SAC). Spatial Audio Coding is a multi-channel compression technique incorporating a backwards compatible downmix, and exploiting the perceptual inter-channel redundancy in multi-channel audio signals to achieve higher compression rates. In order to efficiently represent the spatial image of the multi-channel signal, the MPEG Surround encoder employs a number of spatial parameters:

- Channel level Differences (CLD) describing level differences between two channels,
- Inter channel Correlation / Coherences (ICC) describing the amount of correlation or coherence between two channels and
- Channel Prediction Coefficients (CPC) enabling the recreation of a third channel out of two channels by means of prediction.

During encoding, spatial cues are extracted from the multi-channel audio signal and a downmix is generated. Any number of channels can be used for the downmix, provided that it is less than that used for the original audio signal. The MPEG Surround system supports multi-channel signals of up to 27 channels.

All parameters are efficiently quantized and coded into the spatial bitstream. The quantization process is flexible to allow for a gradual quality versus bit-rate trade-off. For the extreme case where no spatial parameters are employed, the spatial decoder supports an Enhanced Matrix Mode, in which the parameters are derived directly from the down-mix. In addition, support for residual coding is included, that allows coding of the difference between the multi-channel signal as reproduced by the parametric model and the original. Hence, the MPEG Surround system, includes several tools to scale gracefully on the rate/distortion curve, as outlined by the following figure.

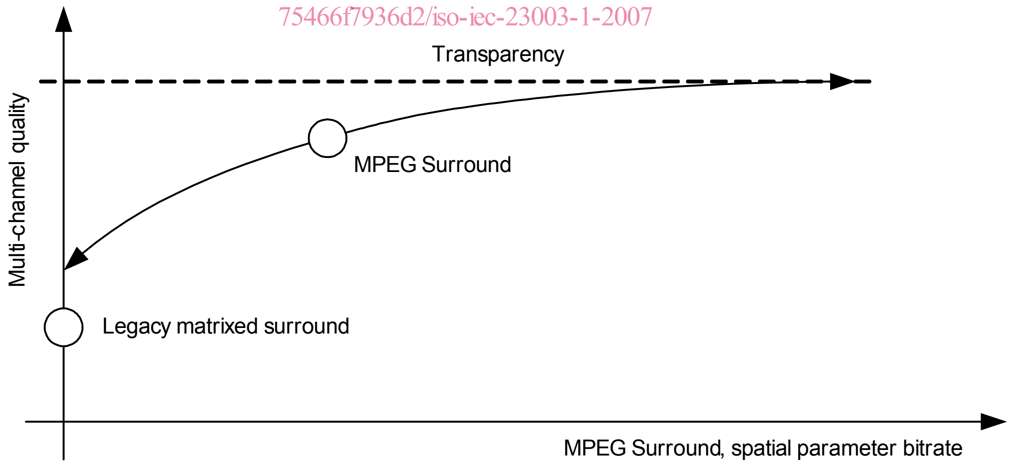

Figure 2 — MPEG Surround bit-rate quality scaling

The downmix can be compressed and transmitted without the need to update existing coders and infrastructures. The spatial cues, or spatial side information, are transmitted in a low bitrate side channel, e.g. the ancillary data portion of the downmix bit-stream.

