
**Information technology — Multimedia
content description interface —**

**Part 4:
Audio**

AMENDMENT 1: Audio extensions

*Technologies de l'information — Interface de description du contenu
multimédia —*

Partie 4: Audio

AMENDEMENT 1: Extensions audio



Reference number
ISO/IEC 15938-4:2002/Amd.1:2004(E)

AVA Documents
Document Sharing Hub
© ISO/IEC 2004

PDF disclaimer

This PDF file may contain embedded typefaces. In accordance with Adobe's licensing policy, this file may be printed or viewed but shall not be edited unless the typefaces which are embedded are licensed to and installed on the computer performing the editing. In downloading this file, parties accept therein the responsibility of not infringing Adobe's licensing policy. The ISO Central Secretariat accepts no liability in this area.

Adobe is a trademark of Adobe Systems Incorporated.

Details of the software products used to create this PDF file can be found in the General Info relative to the file; the PDF-creation parameters were optimized for printing. Every care has been taken to ensure that the file is suitable for use by ISO member bodies. In the unlikely event that a problem relating to it is found, please inform the Central Secretariat at the address given below.

© ISO/IEC 2004

All rights reserved. Unless otherwise specified, no part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from either ISO at the address below or ISO's member body in the country of the requester.

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

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.

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.

Amendment 1 to ISO/IEC 15938-4:2002 was prepared by Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*.

.....

Information technology — Multimedia content description interface —

Part 4: Audio

AMENDMENT 1: Audio extensions

Add at the end of subclause 4.2:

4.3 Handling of multi-channel signals

Introduction:

The framework to handle multi-channel signals is given by the `AudioD` and `AudioDS` Types defined in ISO/IEC 15938-5/Amd.1 (MDS). The new additional attribute `channels` gives the channel numbers that are described by the assigned Descriptor or Description Scheme. However, to prevent some misunderstanding, a more detailed description and handling policy is given in this part of ISO/IEC 15938. In particular a recommendation is given to handle typical surround formats, when only tag names like L, C, R, LS, LR, LFE are known.

By using the `channels` attribute, defined in ISO/IEC 15938-5/Amd.1 (MDS) it is possible to specify which channels should be used for e.g. computing the mean with the extraction method. Therefore, the Descriptor and Description Schemes contain information about these channels only. This is useful in order to separate a multi-channel input signal into subgroups that are closely related, e.g. the Left (L), Center (C) and Right (R) signal of a typical surround format. The highest possible channel number is given in the file-format of the audio media file itself. All numbers given in the `channels` attribute higher than the number of channels given by the media file-format should be ignored.

In the case where the numbering of the audio channels is not explicitly given in the file-format (like 5.1 surround signals), the following convention to number the channels is recommended to be used.

When mapping typical surround file-formats, consisting of tags like (L, R, C, LS, RS, LFE), the scheme shown in Figure AMD1-1 should be followed, in order to reduce the ambiguity between scheme and channel number. To define the channel number, the counting should start at an optional center channel and go from left to right, top to bottom and then from front to back (see for example Figure AMD1-2). An optional rear center will get the last channel number for the standard audio channels. The assigned number can be higher if specialised channels are present, like an LFE channel for low frequency effect signals. Two examples are given in Tables 1 and 2. Furthermore, it is recommended that a textual description of the scheme used inside the `AudioSegmentD-Framework` (defined in ISO/IEC 15938-5) is included. An instantiation example is given in ISO/IEC 15938-5/Amd.1 (MDS), subclause 4.2.4.

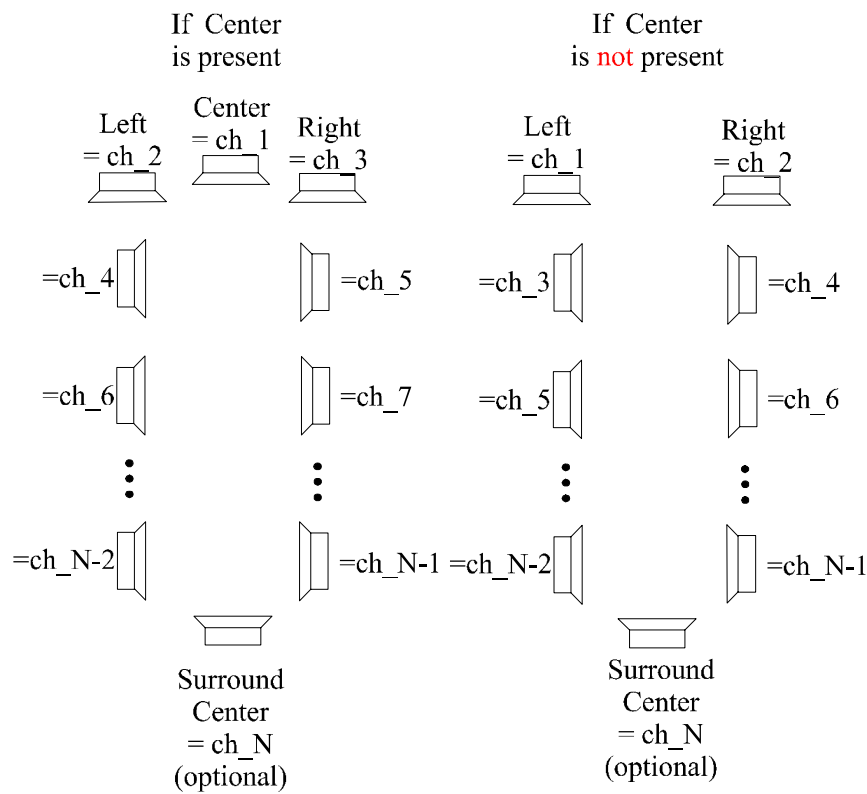


Figure AMD1-1 - Scheme and channel number for typical surround file-formats

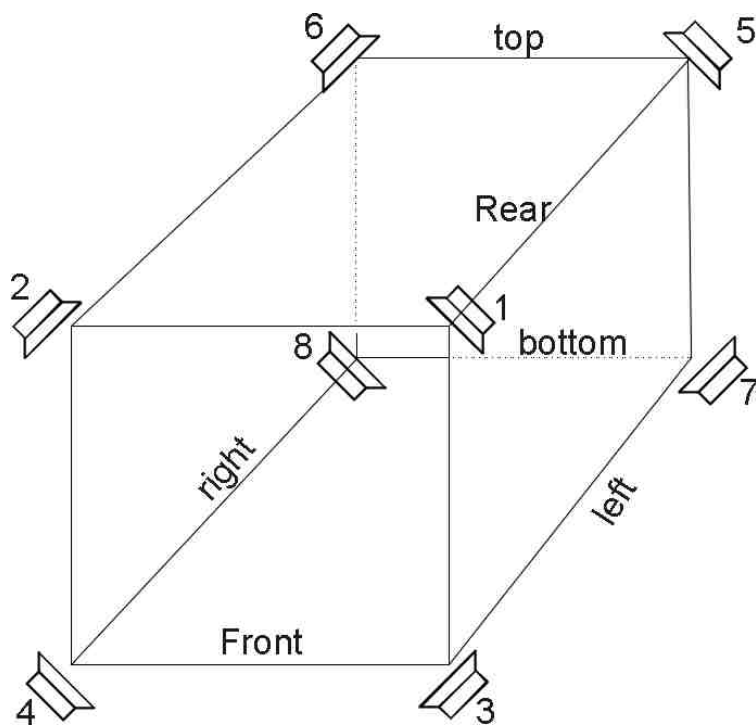


Figure AMD1-2 - Scheme and channel number for a 3D speaker arrangement (example)

Examples for mapping:

Table AMD1-1- Simple Stereo	
Tag name	channel number
Left	1
Right	2

TableAMD1-2- Surround 5.1	
Tag name	Channel number
Center	1
Left	2
Right	3
Left Surround (LS)	4
Right Surround (LR)	5
LFE	6

Replace subclause 6.5.8 by:

6.5.8 WordLexiconType**6.5.8.1 Syntax**

```

<!-- ##### -->
<!-- Definition of WordLexiconType -->
<!-- ##### -->
<complexType name="WordLexiconType">
  <complexContent>
    <extension base="mpeg7:LexiconType">
      <sequence>
        <element name="Token" minOccurs="1" maxOccurs="4294967296">
          <complexType> <!-- New attribute V2 -->
            <simpleContent>
              <extension base="mpeg7:WordType">
                <attribute name="linguisticUnit" use="optional"
                  default="word">
                  <simpleType>
                    <union>
                      <simpleType>
                        <restriction base="NMTOKEN">
                          <enumeration value="word"/>
                          <enumeration value="syllable"/>
                          <enumeration value="morpheme"/>
                          <enumeration value="phrase"/>
                          <enumeration value="component"/>
                          <enumeration value="stem"/>
                          <enumeration value="affix"/>
                          <enumeration value="nonspeech"/>
                          <enumeration value="other"/>
                        </restriction>
                      </simpleType>
                    </union>
                  </simpleType>
                <restriction base="mpeg7:termReferenceType"/>
              </simpleType>
            </simpleContent>
          </complexType>
        </element>
      </sequence>
    </extension>
  </complexContent>
</complexType>

```

```

        </union>
      </simpleType>
    </attribute>
    <attribute name="representation" use="optional"
      default="orthographic">
      <simpleType>
        <restriction base="NMTOKEN">
          <enumeration value="orthographic"/>
          <enumeration value="nonorthographic"/>
        </restriction>
      </simpleType>
    </attribute>
  </extension>
</simpleContent>
</complexType>
</element>
</sequence>
<attribute name="phoneticAlphabet" type="mpeg7:phoneticAlphabetType"
  use="optional"/> <!-- New attribute V2 -->
</extension>
</complexContent>
</complexType>

```

6.5.8.2 Semantics

Name	Definition
WordLexiconType	A list of words (a lexicon). Each entry represents one orthographic representation (spelling) or one non-orthographic representation of a word or linguistic unit. The lexicon is not a phonetic (pronunciation) dictionary.
phoneticAlphabet	The name of the encoding scheme of the phone lexicon. Only needed if phonetic representation is used. See 6.5.9 phoneticAlphabetType
Token	An entry in the lexicon
linguisticUnit	Indicates the type of the linguistic unit that is put into the entry of the word lexicon. The linguistic units are defined as follows. • <i>word</i> – an unit delimited by whitespace. This is the default value. (example: psychoacoustics) • <i>syllable</i> – minimal pronouncable unit (example: psy) • <i>morpheme</i> – minimal meaning bearing unit (example: psycho) • <i>stem</i> – the uninflected base of a word-form, can be polymorphic. (example: psychoacoustic) • <i>affix</i> – needs to be added to a stem to get a word • <i>component</i> – a constituent part of a compound word. Important for compounding languages. (example from German: <i>Forschungs</i> (in English corresponds to "research-")) • <i>nonspeech</i> – noises, both human-produced and background, that are non-linguistic in nature. (example: throat clearing, coughing) • <i>phrase</i> - a sequence of words (e.g. "God bless America") • <i>other</i> - a linguistic unit that does not map onto any of the above Other values that are datatype-valid with respect to mpeg7:termReferenceType are reserved.

representation	Form of representation for a lexicon entry. The kinds of representation are defined as follows. • <i>orthographic</i> – representation of an entry by spelling • <i>nonorthographic</i> – representation of an entry by an identifier that is not synonymous with the spelling of a word. A non-orthographic representation may, for example, encode the phoneme string corresponding to the pronunciation of the entry.
----------------	--

6.5.8.3 Usage, Extraction and Examples (informative)

6.5.8.3.1 Purpose

The word lexicon makes it possible to store the words contained in the lattice. It is common in both speech recognition and spoken document retrieval to include entries in the word lexicon that are “words” only in a wider sense of the term (e.g. acronyms or abbreviations) or not really words at all (e.g. phrases, syllables, morphemes or the individual components of compound words).

The attribute `linguisticUnit` makes it possible to distinguish between these different types of units. Differentiating these units is useful, for instance, when the retrieval algorithm of an application needs to treat different units in different ways. For example stemming, a pre-processing step applied to words, should not be applied to syllables or morphemes. Similarly, different types of units might receive different weightings in the calculation of the retrieval metric.

In some applications, it is also necessary to know if the entry is given in its human-readable form or not. For example, if the entry is human-interpretable and can potentially be displayed to the user or if certain algorithms are applied which are intended for the orthographic form only (e.g. stemming).

6.5.8.3.2 Extraction

The generation of syllable, morpheme, compound and phrase transcriptions of spoken input is performed in the following ways:

- The output of the word or phoneme recognizer is mapped to other linguistic units. For example the recognized word can be transformed into syllables using a syllable generation tool.
- The ASR system produces the desired linguistic unit directly during the recognition process. In this case, the linguistic units are parts of the recognition vocabulary of the speech recognition engine. For example, the dictionary used for the speech recognition system could be composed exclusively of syllables.

6.5.8.3.3 Example

The following example shows a lexicon containing six entries. The first two entries represent syllables and the next two entries represent words. The fifth entry also represents a word, but not in its written form. The last entry represents a phrase.

```
<WordLexicon id="lex1" phoneticAlphabet="other">
  <Token linguisticUnit="syllable" representation="nonorthographic">
    6: n
  </Token>
  <Token linguisticUnit="syllable" representation="nonorthographic">
    6: _s_
  </Token>
  <Token>water</Token>
  <Token linguisticUnit="word">draw</Token>
  <Token representation="nonorthographic">Q e: l e: f a n t</Token>
```

```
<Token linguisticUnit="phrase" representation="orthographic">
  as a rule
</Token>
</WordLexicon>
```

Replace subclause 6.5.12 by:

6.5.12 SpokenContentLinkType

6.5.12.1 Syntax

```
<!-- ##### -->
<!-- Definition of SpokenContentLink datatype -->
<!-- ##### -->
<complexType name="SpokenContentLinkType">
  <attribute name="probability" type="mpeg7:zeroToOneType" default="1.0"/>
  <attribute name="acousticScore" type="mpeg7:nonNegativeReal"
    use="optional" /> <!-- New attribute V2 -->
  <attribute name="nodeOffset" type="mpeg7:unsigned16" default="1"/>
</complexType>
```

6.5.12.2 Semantics

Name	Definition
SpokenContentLinkType	The structure of a word or phone link in the lattice
probability	The probability of this link. In a crude sense, this is to indicate which links are more likely than others, with larger numbers indicating higher likelihood.
nodeOffset	The node to which this link leads, specified as a relative offset and defaulting to 1. A node offset leading out of the current block implicitly refers to the next block. A node offset cannot span a whole block, i.e., a link from a node in block 3 must lead to a node in block 3 or block 4.
acousticScore	The score assigned by the acoustic models of the speech recognition engine only. It is given in logarithmic scale (base e) and indicates the quality of the match between the acoustic models and the corresponding signal segment. A higher value indicates a better match.

Add a new subclause 6.7:

6.7 Audio Signal Quality

6.7.1 Introduction

If an AudioSegment DS contains a piece of music, several features describing the signal's quality can be computed to describe the quality attributes. The AudioSignalQualityType contains these quality attributes and uses the ErrorEventType to handle typical errors that occur in audio data and in the transfer process from analog audio to the digital domain. However, note that this DS is not applicable to describe the subjective sound quality of audio signals resulting from sophisticated digital signal processing, including the use of noise shaping or other techniques based on perceptual/psychoacoustic considerations.

For example, in the case of searching an audio file on the Internet, quality information could be used to determine which one should be downloaded among several search results. Another application area would be an archiving system. There, it would be possible to browse through the archive using quality information, and also the information could be used to decide if a file is of sufficient quality to be used e.g. for broadcasting.

6.7.2 Conventions

The description of the Descriptors refers to the input signal x . If x is a multi channel signal, then the signal for a certain channel is designated as x_n for the n -th channel. The functions $\max()$, $\min()$ and $\text{mean}()$ are used as defined in ISO/IEC CD 15938-4 (Audio Part). The function $\text{abs}()$ calculates the absolute value.

6.7.3 Audio Signal Quality Description Scheme

The `AudioSignalQualityType` is a set of `AudioQuality` Descriptors and some additional tools for handling and describing audio signal quality information. In particular the handling of single error events in audio streams is considered.

6.7.3.1 Syntax

```
<!-- ##### -->
<!-- Definition of AudioSignalQuality DS -->
<!-- ##### -->
<complexType name="AudioSignalQualityType">
  <complexContent>
    <extension base="mpeg7:AudioDSType">
      <sequence>
        <!-- Summaries-->
        <element name="Operator" type="mpeg7:PersonType" minOccurs="0"/>
        <element name="UsedTool" type="mpeg7:CreationToolType" minOccurs="0"/>
        <element name="BackgroundNoiseLevel"
type="mpeg7:BackgroundNoiseLevelType"/>
        <element name="RelativeDelay" type="mpeg7:RelativeDelayType"/>
        <element name="Balance" type="mpeg7:BalanceType"/>
        <element name="DcOffset" type="mpeg7:DcOffsetType"/>
        <element name="CrossChannelCorrelation"
type="mpeg7:CrossChannelCorrelationType"/>
        <element name="Bandwidth" type="mpeg7:BandwidthType"/>
        <!-- High Level Feature-->
        <element name="TransmissionTechnology"
type="mpeg7:TransmissionTechnologyType" minOccurs="0"/>
        <!-- Events-->
        <element name="ErrorEventList" minOccurs="0">
          <complexType>
            <sequence>
              <element name="ErrorEvent" type="mpeg7:ErrorEventType"
minOccurs="0" maxOccurs="unbounded"/>
            </sequence>
          </complexType>
        </element>
      </sequence>
      <!-- Attributes-->
      <attribute name="IsOriginalMono" type="boolean"/>
      <attribute name="BroadcastReady" type="boolean" use="optional"/>
    </extension>
  </complexContent>
</complexType>
```