Network Working Group
Request for Comments: 5691
Updates: 3640
Category: Standards Track

F. de Bont
Philips Electronics
S. Doehla
Fraunhofer IIS
M. Schmidt
Dolby Laboratories
R. Sperschneider
Fraunhofer IIS

October 2009

RTP Payload Format for Elementary Streams with MPEG Surround Multi-Channel Audio

Abstract

This memo describes extensions for the RTP payload format defined in RFC 3640 for the transport of MPEG Surround multi-channel audio. Additional Media Type parameters are defined to signal backwards-compatible transmission inside an MPEG-4 Audio elementary stream. In addition, a layered transmission scheme that doesn't use the MPEG-4 systems framework is presented to transport an MPEG Surround elementary stream via RTP in parallel with an RTP stream containing the downmixed audio data.

Status of This Memo

This document specifies an Internet standards track protocol for the Internet community, and requests discussion and suggestions for improvements. Please refer to the current edition of the "Internet Official Protocol Standards" (STD 1) for the standardization state and status of this protocol. Distribution of this memo is unlimited.

Copyright Notice

Copyright (c) 2009 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the BSD License.

Table of Contents

1. Introduction 2
2. Conventions 3
3. Definitions and Abbreviations 3
3.1. Definitions 3
3.2. Abbreviations 4
4. Transport of MPEG Surround 4
4.1. Embedded Spatial Audio Data in AAC Payloads 4
4.2. MPEG Surround Elementary Stream 5
4.2.1. Low Bitrate MPEG Surround 7
4.2.2. High Bitrate MPEG Surround 8
5. IANA Considerations 8
5.1. Media Type Registration 9
5.2. Registration of Mode Definitions with IANA 9
5.3. Usage of SDP 10
6. Security Considerations 10
7. References 11
7.1. Normative References 11
7.2. Informative References 11

1. Introduction

MPEG Surround (Spatial Audio Coding, SAC) [23003-1] is an International Standard that was finalized by MPEG in January 2007. It 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 significantly lower data rate than the data rate required for transmitting all N channels, making the coding very efficient while at the same time ensuring compatibility with M channel devices.

The MPEG Surround standard incorporates a number of tools that enable features that allow for broad application of the 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 decoder input can be made compatible to existing matrixed surround technologies.

As an example, for 5.1 multi-channel audio, the MPEG Surround encoder creates a stereo (or mono) downmix signal and spatial information describing the full 5.1 material in a highly efficient, parameterised format. The spatial information is transmitted alongside the downmix.

By using MPEG Surround, existing services can easily be upgraded to provide surround sound in a backwards-compatible fashion. While a stereo decoder in an existing legacy consumer device ignores the MPEG Surround data and plays back the stereo signal without any quality degradation, an MPEG-Surround-enabled decoder will deliver high quality, multi-channel audio.

The MPEG Surround decoder can operate in modes that render the multi-channel signal to multi-channel or stereo output, or it can operate in a two-channel headphone mode to produce a virtual surround output signal.

2. Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

3. Definitions and Abbreviations

3.1. Definitions

This memo makes use of the definitions specified in [14496-1], [14496-3], [23003-1], and [RFC3640]. Frequently used terms are summed up for convenience:

Access Unit: An MPEG Access Unit is the smallest data entity to which timing information is attributed. In the case of audio, an Access Unit is the smallest individually accessible portion of coded audio data within an elementary stream.

AudioSpecificConfig(): Extends the class DecoderSpecificInfo(), as defined in [14496-1], when the objectType indication refers to a stream complying with [14496-3]. AudioSpecificConfig() is used as the configuration structure for MPEG-4 audio as specified in [14496-3]. It contains the field audioObjectType, which distinguishes between the different audio codecs defined in [14496-3], general audio information (e.g., the sampling frequency and number of channels), and further codec-dependent information structures.

SpatialSpecificConfig(): Configuration structure for MPEG Surround audio coding, as specified in [23003-1]. An AudioSpecificConfig() with an audioObjectType of value 30 contains a SpatialSpecificConfig() structure.

3.2. Abbreviations

AOT: Audio Object Type
AAC: Advanced Audio Coding
ASC: AudioSpecificConfig() structure
AU: Access Unit
HE AAC: High Efficiency AAC
PLI: Profile and Level Indication
SSC: SpatialSpecificConfig() structure

4. Transport of MPEG Surround

From a top-level perspective, MPEG Surround data can be subdivided into configuration data contained in the SpatialSpecificConfig() (SSC) and the SpatialFrame(), which contains the MPEG Surround payload. The configuration data can be signaled in-band or out-of-band. In the case of in-band signaling the SSC is conveyed in a SacDataFrame() jointly with a SpatialFrame(). In the case of out-of-band signaling, the SSC is transmitted to the decoder separately, e.g., by Session Description Protocol (SDP) [RFC4566] means.

SpatialFrame()s may be transmitted either embedded into the downmix stream (Section 4.1) or as individual elementary streams besides the downmix audio stream (Section 4.2).

The buffer definition for AAC decoders limits the size of an AU, as specified in [14496-3]. For high-bitrate applications that exceed this limit, all MPEG Surround data MUST be put in a separate stream, as defined in Section 4.2.

4.1. Embedded Spatial Audio Data in AAC Payloads

[14496-3] defines the extension_payload() as a mechanism for transport of extension data inside AAC payloads. Typical extension data include Spectral Band Replication (SBR) data and MPEG Surround data, i.e., a SacDataFrame() in extension_payload()s of type EXT_SAC_DATA. extension_payload()s reside inside the downmix AAC elementary stream. The resulting single elementary stream is transported as specified in [RFC3640]. As AAC decoders are required to skip unknown extension data, MPEG Surround data can be embedded in backwards-compatible fashion and be transported with the mechanism already described in [RFC3640].

The SacDataFrame() includes a SpatialFrame() and an optional header that contains an SSC. Any SSC in a SacDataFrame() MUST be identical to the SSC conveyed via SDP for that stream.

No new mode is introduced for SpatialFrame()s being embedded into AAC payloads. Either the mode AAC-lbr or the mode AAC-hbr SHOULD be used. The additional Media Type parameters, as defined in Section 5.1, SHOULD be present when SpatialFrame()s are embedded into AAC payloads.

For example:

```
m=audio 5000 RTP/AVP 96
a=rtpmap:96 mpeg4-generic/48000/2
a=fmtp:96 streamType=5; profile-level-id=44; mode=AAC-hbr; config=131
  056E598; sizeLength=13; indexLength=3; indexDeltaLength=3; constant
  Duration=2048; MPS-profile-level-id=55; MPS-config=F1B4CF920442029B
  501185B6DA00;
```

In this example, the stream specifies the HE AAC Profile at Level 2 [Profile and Level Indication (PLI) 44] and the config string contains the hexadecimal representation of the HE AAC ASC [audioObjectType=2 (AAC LC); extensionAudioObjectType=5 (SBR); samplingFrequencyIndex=0x6 (24kHz); extensionSamplingFrequencyIndex=0x3 (48kHz); channelConfiguration=2 (2.0 channels)] of the downmix AAC elementary stream that is using explicit backwards-compatible signaling.

Furthermore, the stream specifies the MPEG Surround Baseline Profile at Level 3 (PLI55) and the MPS-config string contains the hexadecimal representation of the MPEG Surround ASC [audioObjectType=30 (MPEG Surround); samplingFrequencyIndex=0x3 (48kHz); channelConfiguration=6 (5.1 channels); sacPayloadEmbedding=1; SSC=(48 kHz; 32 slots; 525 tree; ResCoding=1; ResBands=[0,13,13,13])].

Note that the a=fmtp line of the example above has been wrapped to fit the page; it would comprise a single line in the SDP file.

4.2. MPEG Surround Elementary Stream

MPEG Surround SpatialFrame()s can be present in an individual elementary stream. This stream complements the stream containing the downmix audio data, which may be coded by an arbitrary coding scheme. MPEG Surround elementary streams are packetized as specified in [RFC3640]. The mode signaled and used for an MPEG Surround elementary stream MUST be either MPS-hbr or MPS-lbr. The MPS-hbr mode SHALL be used when the frame size may exceed 63 bytes, e.g., when high-bitrate residual coding is in use.

The dependency relationships between the MPEG Surround elementary stream and the downmix stream are signaled as specified in [RFC5583].

The media clocks of the MPEG Surround elementary stream and the downmix stream SHALL operate in the same clock domain, i.e., the clocks are derived from a common clock and MUST NOT drift. RTCP sender reports MUST indicate that the stream timestamps are not drifting, i.e., that a single sender report for each stream is sufficient to establish unambiguous timing. The sampling rate of the MPEG Surround signal and the decoded downmix signal MUST be identical.

If HE AAC is used as the coding scheme for the downmix, the RTP clock-rate of the downmix MAY be the sampling rate of the AAC core, i.e., the clock-rate of the MPEG Surround elementary stream is an integer multiple of the clock-rate of the downmix stream.

Note that separate RTP streams have different random RTP timestamp offsets, and therefore RTCP MUST be used to synchronize the coded downmix audio data and the MPEG Surround elementary stream.

For example:

```
a=group:DDP L1 L2

m=audio 5000 RTP/AVP 96
a=rtpmap:96 mpeg4-generic/48000/2
a=fmtp:96 streamType=5; profile-level-id=44; mode=AAC-hbr; config=2B1
  18800; sizeLength=13; indexLength=3; indexDeltaLength=3; constantDu
  ration=2048
a=mid:L1

m=audio 5002 RTP/AVP 97
a=rtpmap:97 mpeg4-generic/48000/6
a=fmtp:97 streamType=5; profile-level-id=55; mode=MPS-hbr; config=F1B
  0CF920460029B601189E79E70; sizeLength=13; indexLength=3;  indexDelt
  aLength=3; constantDuration=2048
a=mid:L2
a=depend:97 lay L1:96
```

In this example, the first stream specifies the HE AAC Profile at Level 2 (PLI44) and the config string contains the hexadecimal representation of the HE AAC ASC [audioObjectType=2 (AAC LC); extensionAudioObjectType=5 (SBR); samplingFrequencyIndex=0x6 (24kHz); extensionSamplingFrequencyIndex=0x3 (48kHz); channelConfiguration=2 (2.0 channels)].

The second stream specifies Baseline MPEG Surround Profile at Level 3 (PLI55) and the config string contains the hexadecimal representation of the ASC [AOT=30(MPEG Surround); 48 kHz; 5.1-ch; sacPayloadEmbedding=0; SSC=(48 kHz; 32 slots; 525 tree; ResCoding=1; ResBands=[7,7,7,7])].

Note that the a=fmtp lines of the example above have been wrapped to fit the page; they would each comprise a single line in the SDP file.

4.2.1. Low Bitrate MPEG Surround

This mode is signaled by mode=MPS-lbr. This mode supports the transport of one or more complete Access Units, each consisting of a single MPEG Surround SpatialFrame(). The AUs can be variably sized and interleaved. The maximum size of a SpatialFrame() is 63 bytes. Fragmentation MUST NOT be used in this mode. Receivers MUST support de-interleaving.

The payload configuration is the same as in the AAC-lbr mode. It consists of the AU Header Section, followed by concatenated AUs. Note that Access Units are byte-aligned. The Auxiliary Section MUST be empty in the MPS-lbr mode. The 1-octet AU-header MUST provide:

1. the size of each AAC frame, encoded as 6 bits.

2. 2 bits of index information for computing the sequence (and hence timing) of each SpatialFrame().

The concatenated AU Header Section MUST be preceded by the 16-bit AU-headers-length field.

In addition to the required Media format parameters, the following parameters MUST be present with fixed values: sizeLength (fixed value 6), indexLength (fixed value 2), and indexDeltaLength (fixed value 2). The parameter maxDisplacement MUST be present when interleaving. SpatialFrame()s always have a fixed duration per AU; the fixed duration MUST be signaled by the Media format parameter constantDuration.

The value of the "config" parameter is the hexadecimal representation of the ASC, as defined in [14496-3], with an AOT of 30 and the sacPayloadEmbedding flag set to 0.

The "profile-level-id" parameter SHALL contain a valid PLI for MPEG Surround, as specified in [14496-3].

4.2.2. High Bitrate MPEG Surround

This mode is signaled by mode=MPS-hbr. This mode supports the transportation of either one fragment of an Access Unit or one complete AU or several complete AUs. Each AU consists of a single MPEG Surround SpatialFrame(). The AUs can be variably sized and interleaved. The maximum size of a SpatialFrame() is 8191 bytes. Receivers MUST support de-interleaving.

The payload configuration is the same as in the AAC-hbr mode. It consists of the AU Header Section, followed by either one SpatialFrame(), a fragment of a SpatialFrame(), or several concatenated SpatialFrame()s. Note that Access Units are byte-aligned. The Auxiliary Section MUST be empty in the MPS-hbr mode. The 2-octet AU-header MUST provide:

1. the size of each AAC frame, encoded as 13 bits.

2. 3 bits of index information for computing the sequence (and hence timing) of each SpatialFrame(), i.e., the AU-Index or AU-Index-delta field.

Each AU-Index field MUST be coded with the value 0. The concatenated AU Header Section MUST be preceded by the 16-bit AU-headers-length field.

In addition to the required Media format parameters, the following parameters MUST be present with fixed values: sizeLength (fixed value 13), indexLength (fixed value 3), and indexDeltaLength (fixed value 3). The parameter maxDisplacement MUST be present when interleaving. SpatialFrame()s always have a fixed duration per AU; the fixed duration MUST be signaled by the Media format parameter constantDuration.

The value of the "config" parameter is the hexadecimal representation of the ASC, as defined in [14496-3], with an AOT of 30 and the sacPayloadEmbedding flag set to 0.

The "profile-level-id" parameter SHALL contain a valid PLI for MPEG Surround, as specified in [14496-3].

5. IANA Considerations

This memo defines additional optional format parameters to the Media type "audio" and its subtype "mpeg4-generic". These parameters SHALL only be used in combination with the AAC-lbr or AAC-hbr modes (cf. Section 3.3 of [RFC3640]) of "mpeg4-generic".

5.1. Media Type Registration

This memo defines the following additional optional parameters, which SHALL be used if MPEG Surround data is present inside the payload of an AAC elementary stream.

MPS-profile-level-id: A decimal representation of the MPEG Surround Profile and Level indication as defined in [14496-3]. This parameter MUST be used in the capability exchange or session set-up procedure to indicate the MPEG Surround Profile and Level that the decoder must be capable of in order to decode the stream.

MPS-config: A hexadecimal representation of an octet string that expresses the AudioSpecificConfig (ASC), as defined in [14496-3], for MPEG Surround. The ASC is mapped onto the hexadecimal octet string in a most significant bit (MSB)-first basis. The AOT in this ASC SHALL have the value 30. The SSC inside the ASC MUST have the sacPayloadEmbedding flag set to 1.

5.2. Registration of Mode Definitions with IANA

This section of this memo requests the registration of the "MPS-hbr" value and the "MPS-lbr" value for the "mode" parameter of the "mpeg4-generic" media subtype within the media type "audio". The "mpeg4-generic" media subtype is defined in [RFC3640], and [RFC3640] defines a repository for the "mode" parameter. This memo registers the modes "MPS-hbr" and "MPS-lbr" to support MPEG Surround elementary streams.

Media type name:

audio

Subtype name:

mpeg4-generic

Required parameters:

The "mode" parameter is required by [RFC3640]. This memo specifies the additional modes "MPS-hbr" and "MPS-lbr", in accordance with [RFC3640].

Optional parameters:

For the modes "AAC-hbr" and "AAC-lbr", this memo specifies the additional optional parameters "MPS-profile-level-id" and "MPS-config". See Section 4.1 for usage details.

Optional parameters for the modes "MPS-hbr" and "MPS-lbr" may be used as specified in [RFC3640]. The optional parameters "MPS-profile-level-id" and "MPS-config" SHALL NOT be used for the modes "MPS-hbr" and "MPS-lbr".

5.3. Usage of SDP

It is assumed that the Media format parameters are conveyed via an SDP message, as specified in Section 4.4 of [RFC3640].

6. Security Considerations

RTP packets using the payload format defined in this specification are subject to the security considerations discussed in the RTP specification [RFC3550], in the RTP payload format specification for MPEG-4 elementary streams [RFC3640] (which is extended with this memo), and in any applicable RTP profile. The main security considerations for the RTP packet carrying the RTP payload format defined within this memo are confidentiality, integrity, and source authenticity. Confidentiality is achieved by encryption of the RTP payload. Integrity of the RTP packets is achieved through a suitable cryptographic integrity-protection mechanism. Such a cryptographic system may also allow the authentication of the source of the payload. A suitable security mechanism for this RTP payload format should provide confidentiality, integrity protection, and source authentication capable of at least determining if an RTP packet is from a member of the RTP session.

The AAC audio codec includes an extension mechanism to transmit extra data within a stream that is gracefully skipped by decoders that do not support this extra data. This covert channel may be used to transmit unauthorized data in an otherwise valid stream.

Note that the appropriate mechanism to provide security to RTP and payloads following this memo may vary. It is dependent on the application, the transport, and the signaling protocol employed. Therefore, a single mechanism is not sufficient; although, if suitable, usage of the Secure Real-time Transport Protocol (SRTP) [RFC3711] is recommended. Other mechanisms that may be used are IPsec [RFC4301] and Transport Layer Security (TLS) [RFC5246] (RTP over TCP); other alternatives may exist.

7. References

7.1. Normative References

[14496-1] MPEG, "ISO/IEC International Standard 14496-1 - Coding of audio-visual objects, Part 1 Systems", 2004.

[14496-3] MPEG, "ISO/IEC International Standard 14496-3 - Coding of audio-visual objects, Part 3 Audio", 2009.

[23003-1] MPEG, "ISO/IEC International Standard 23003-1 - MPEG Surround (MPEG D)", 2007.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, July 2003.

[RFC3640] van der Meer, J., Mackie, D., Swaminathan, V., Singer, D., and P. Gentric, "RTP Payload Format for Transport of MPEG-4 Elementary Streams", RFC 3640, November 2003.

[RFC4566] Handley, M., Jacobson, V., and C. Perkins, "SDP: Session Description Protocol", RFC 4566, July 2006.

[RFC5583] Schierl, T. and S. Wenger, "Signaling Media Decoding Dependency in the Session Description Protocol (SDP)", RFC 5583, July 2009.

7.2. Informative References

[RFC3711] Baugher, M., McGrew, D., Naslund, M., Carrara, E., and K. Norrman, "The Secure Real-time Transport Protocol (SRTP)", RFC 3711, March 2004.

[RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, December 2005.

[RFC5246] Dierks, T. and E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.2", RFC 5246, August 2008.

Authors' Addresses

Frans de Bont
Philips Electronics
High Tech Campus 5
5656 AE Eindhoven,
NL

Phone: ++31 40 2740234
EMail: frans.de.bont@philips.com

Stefan Doehla
Fraunhofer IIS
Am Wolfmantel 33
91058 Erlangen,
DE

Phone: +49 9131 776 6042
EMail: stefan.doehla@iis.fraunhofer.de

Malte Schmidt
Dolby Laboratories
Deutschherrnstr. 15-19
90537 Nuernberg,
DE

Phone: +49 911 928 91 42
EMail: malte.schmidt@dolby.com

Ralph Sperschneider
Fraunhofer IIS
Am Wolfmantel 33
91058 Erlangen,
DE

Phone: +49 9131 776 6167
EMail: ralph.sperschneider@iis.fraunhofer.de