Sign Language Video Encoding for Digital Cinema

ISDCF Document 13

July 18, 2018

Version 1.0

ISDCF provides this document for convenience only to be used voluntarily at your own risk, no warranty expressed or implied.
Visit isdcf.com for more information.

1. Introduction

This document describes a method for the encoding and packaging of a sign language video track within a Digital Cinema Package (DCP) for distribution to exhibition.

This scheme benefits from the following advantages:

- compatibility with all existing Digital Cinema projection systems
- supports random access playback
- synchronization with audio at the output of the media block

Decoding of the video track may be performed internal to or external to the Media Block.

2. Video Format

This section describes the format requirement for the sign language video track.

2.1. Codec

The video codec shall be VP9[1].

2.2. Resolution

The video shall be 480 pixels wide and 640 pixels high (*i.e.*, portrait orientation).

2.3. Framerate

The video frame rate shall be 24.0 fps regardless of any other DCP frame rates in use.

2.4. Bitrate

The VP9 bitstream shall have a maximum bitrate of 1.0 Mbps.

2.5. Colorspace

The video shall be encoded as Y'UV.

2.6. Pixel Format

The video shall use Y'UV420p chroma sub-sampling.

[1] VP9 is a high-quality, open video format; see https://www.webmproject.org/vp9/.

3. Mapping Into the DCP

This section describes the method for mapping the above video bitstream into a digital audio channel for inclusion in a DCP.

3.1. Background on Digital Cinema Audio

DCPs carry audio as a sequence of uncompressed frames of (up to) 16 channels of 24-bit Pulse-Code Modulation (PCM)[2]. The PCM has a sample rate of 48 kHz. Note that 96 kHz is supported in digital cinema but shall be prohibited for this application. Each audio frame has a duration:

$$\text{duration} = 1/\text{e}$$

where e is Edit Rate of the composition.

During playback, projection systems sequentially decrypt each frame of audio and output each of its channels to their corresponding AES/EBU digital output to form 16 independent digital audio streams. Each of these 16 digital streams is therefore operating at the following *fixed bitrate*:

$$48000 \text{ samples/s} * 24 \text{ bit/sample} = 1.152 \text{ Mb/s}.$$

3.2. VP9 Chunking

VP9 is inherently a variable bitrate codec. To allow for carriage in fixed bitrate digital audio, and to allow for operator-initiated timeline jumps (*i.e.*, "trick play"), the video must be encoded into discrete chunks that get distributed evenly throughout the digital audio program. Each chunk contains both an EBML Header and VP9 Segment.

The duration of each VP9 chunk shall be:

$$D_c = \mathbf{2\ seconds}.$$

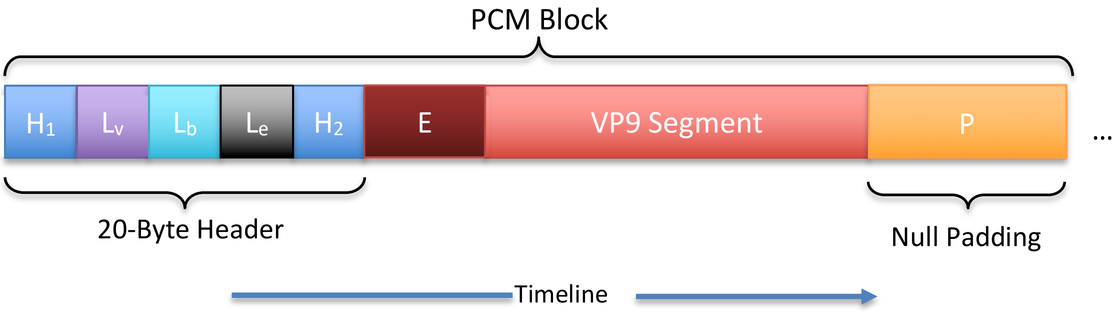
3.3. PCM Block Structure

Each segment of VP9 is carried in a block corresponding to PCM essence of equal duration. Thus, the length of each PCM block is:

$$\mathbf{L_b} = 48{,}000 \text{ samples/s} \cdot 3 \text{ bytes/sample} \cdot D_c \mathbf{= 288{,}000\ bytes}$$

A PCM block is composed of a 20-byte header, followed by the VP9 EBML Header, followed by the VP9 segment, followed by zero or more null bytes. See diagram below for details:

[2] See EBU Tech 3285.

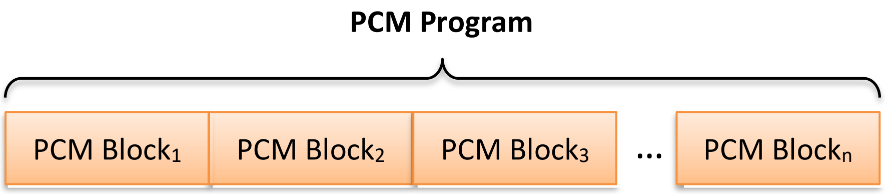


Where:

H_1 = 0xFFFFFFFF
L_v = Length of VP9 segment in bytes (32-bit unsigned integer, big-endian)
L_b = Length of PCM Block in bytes (32-bit unsigned integer, big-endian)
L_e = Length of VP9 EBML Header in bytes (32-bit unsigned integer, big-endian)
H_2 = 0xFFFFFFFF
E **=** VP9 EBML Header
P = A sequence of $\mathbf{L_b - L_v - L_e - 20}$ null bytes

3.4. Complete PCM Program

Each PCM Block described above is combined to form the final PCM program as follows:



3.5. Audio Channel Number

The sign language PCM program shall reside in channel 15 of a 16-channel DCP Main Sound track file.

4. Forensic Marking

To prevent the corruption of an encrypted sign language track by the audio forensic marking feature of the digital cinema projection system, associated Key Delivery Messages (KDMs):

a) shall carry the selective audio FM mark flag as specified at Section 9.4.6.2.3(d) of the DCI System Specification, set to a value smaller than or equal to 14 (see Note below); and

b) may also carry the http://www.smpte-ra.org/430-1/2006/KDM#mrkflg-audio-disable flag for compatibility with legacy image media blocks (IMBs).

ISDCF provides this document for convenience only to be used voluntarily at your own risk, no warranty expressed or implied.
Visit isdcf.com for more information.

Note: The value of the selective audio FM mark flag is based on the specific sound channels that cannot be forensically-marked. In particular, when targeting a Composition containing Motion Data and External Sync Signal in addition to the Sign Language Video, the selective audio FM mark flag is set to a value smaller than or equal to 12.

5. SMPTE DCP Mastering

These guidelines apply to SMPTE DCPs that carry Sign Language Video.

5.1. Composition Playlist Metadata

SMPTE Composition Playlists (CPLs) that carry a Sign Language Video track should indicate the presence of this track using CPL Metadata[3]. The following extension element should be used within the `ExtensionMetadataList` element:

```
<ExtensionMetadata scope="http://isdcf.com/2017/10/SignLanguageVideo">
  <Name>Sign Language Video</Name>
  <PropertyList>
    <Property>
      <Name>Language Tag</Name>
      <Value>DESCRIPTION</Value>
    </Property>
  </PropertyList>
</ExtensionMetadata>
```

Where ***DESCRIPTION*** is a Language-Tag, as specified in IETF RFC 5646, that identifies the sign language present in the Sign Language Video Track.

The IANA Language Subtag Registry[4] lists subtags for sign languages.

5.2. Track files using MXF Multichannel Audio (MCA) Framework

The items of the Audio Channel Label Subdescriptor associated with a sign language video signal shall be set according to the following table:

MCA Label Dictionary ID	MCA Tag Name	MCA Tag Symbol	RFC 5646 Spoken Language
06.0E.2B.34.04.01.01.0D.0D.0F.03.02.01.01.00.00	Sign Language Video Stream	SLVS	Same as the Soundfield Group Label Sub Descriptor

6. Decoder Behavior

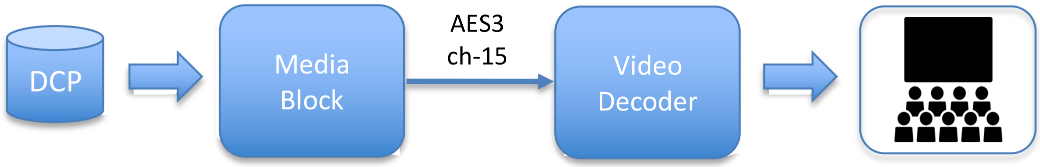
This section defines the role of a decoder as well as its basic behavior.

[3] See SMPTE ST 429-16.

[4] See https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry.

6.1. Overview

A decoder is a device that is connected to a Digital Cinema Media Block such that it can consume the AES3 stream being output on channel 15 as depicted below:



During playback, the decoder processes this AES3 stream in order to output native video to the auditorium.

Note: The method used to distribute real-time video from the decoder to the auditorium is outside the scope of this document.

6.2. Robust Block Detection

Decoders should not expect that all data presented to it on channel 15 conforms to the block structure outlined in Section 3.3. Not all compositions that contain essence on channel 15 will carry video as defined in this document. Some may contain silence, or other audio essence. Additionally, those compositions that do carry sign language video, may not carry valid blocks throughout the entire timeline. Therefore, a robust parser should continually scan its incoming stream for data that resembles a valid header, and data that does not resemble a valid header should be ignored.

ISDCF provides this document for convenience only to be used voluntarily at your own risk, no warranty expressed or implied.
Visit isdcf.com for more information.