SMPTE STANDARD

Immersive Audio Metadata

Table of Contents

Approved June 18. 2018

Foreword

SMPTE (the Society of Motion Picture and Television Engineers) is an internationally recognized standards developing organization. Headquartered and incorporated in the United States of America, SMPTE has members in over 80 countries on six continents. SMPTE's Engineering Documents, including Standards, Recommended Practices, and Engineering Guidelines, are prepared by SMPTE's Technology Committees. Participation in these Committees is open to all with a bona fide interest in their work. SMPTE cooperates closely with other standards-developing organizations, including ISO, IEC and ITU.

SMPTE Engineering Documents are drafted in accordance with the rules given in Part XIII of its Administrative Practices. The Technology Committee 25CSS Working Group on Interoperability of Immersive Sound in Digital Cinema prepared this SMPTE Engineering Document.

Intellectual Property

At the time of publication no notice had been received by SMPTE claiming patent rights essential to the implementation of this Engineering Document. However, attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. SMPTE shall not be held responsible for identifying any or all such patent rights.

Introduction

Immersive audio consists of Audio Channels and/or Audio Objects, which can be utilized by the content creator to design a sound track with sounds above and around the listener. Immersive audio combines metadata with audio essence, which allows the Audio Objects and Audio Channels in the sound track to be rendered successfully into multiple Loudspeaker configurations. This document identifies metadata required for immersive audio for cinema.

1 Scope

This standard defines metadata for use in creating immersive audio content for cinema. The standard defines the metadata items that shall be supported in immersive audio content and if appropriate, the range of values, value precisions, and cardinal values. The format of the metadata items when carried in a bitstream or file is outside the scope of this specification and is provided in a separate document.

2 Conformance Notation

Normative text is text that describes elements of the design that are indispensable or contains the conformance language keywords: "shall", "should", or "may". Informative text is text that is potentially helpful to the user, but not indispensable, and can be removed, changed, or added editorially without affecting interoperability. Informative text does not contain any conformance keywords.

All text in this document is, by default, normative, except: the Introduction, any section explicitly labeled as "Informative" or individual paragraphs that start with "Note:"

The keywords "shall" and "shall not" indicate requirements strictly to be followed in order to conform to the document and from which no deviation is permitted.

The keywords, "should" and "should not" indicate that, among several possibilities, one is recommended as particularly suitable, without mentioning or excluding others; or that a certain course of action is preferred but not necessarily required; or that (in the negative form) a certain possibility or course of action is deprecated but not prohibited.

The keywords "may" and "need not" indicate courses of action permissible within the limits of the document.

The keyword “reserved” indicates a provision that is not defined at this time, shall not be used, and may be defined in the future. The keyword “forbidden” indicates “reserved” and in addition indicates that the provision will never be defined in the future.

A conformant implementation according to this document is one that includes all mandatory provisions ("shall") and, if implemented, all recommended provisions ("should") as described. A conformant implementation need not implement optional provisions ("may") and need not implement them as described.

Unless otherwise specified, the order of precedence of the types of normative information in this document shall be as follows: Normative prose shall be the authoritative definition; Tables shall be next; followed by formal languages; then figures; and then any other language forms.

3 Normative References

The following standards contain provisions that, through reference in this text, constitute provisions of this Standard. At the time of publication, the editions indicated were valid. All standards are subject to revision, and parties to agreements based on this Standard are encouraged to investigate the possibility of applying the most recent edition of the standards indicated below.

SMPTE ST 377-4:2012 MXF Multichannel Audio Labeling Framework

SMPTE ST 428-12:2013 D-Cinema Distribution Master – Common Audio Channels and Soundfield Groups

SMPTE ST 2098-5:2018 D-Cinema Immersive Audio Channels and Soundfield Groups

4 Terms and Definitions

4.1 Audio Object

A segment of audio essence with associated metadata describing positional and other properties which may vary with time.

4.2 Bed

A Soundfield Group, such as a 5.1, 7.1 or 9.1, that is typically present for the duration of the program and serves as the foundation of the immersive soundtrack mix.

4.3 Group

A set of Audio Channels, Audio Objects or components that belong together (e.g. stereo, 5.1).

4.4 Renderer

A device or algorithm that reads in audio tracks and associated metadata and converts them to another set of audio signals destined for individual reproduction devices.

4.5 Sample

A discrete number representing the amplitude of an audio signal at an instant in time.

4.6 Target Environment

A specific set of conditions that is present in the playback environment.

4.7 Zone

A defined region within a generic listening environment.

5 Channel Metadata

5.1 General

The following subsections specify metadata requirements associated with Audio Channels.

5.2 Channel Identifier

This Identifier uniquely identifies an Audio Channel. No two Audio Channels shall have the same Channel Identifier at any instant in time.

5.3 Routing Destination

This metadata item identifies the single Loudspeaker or other reproduction device associated with the channel. For purposes of this document, “other reproduction device” may include an array of Loudspeakers driven in unison, or it may include a process to modify the audio prior to presentation. The metadata shall be coded such that the desired routing destination is unambiguously identified.

5.4 Waveform Reference

This metadata item references audio essence associated with the channel. The reference shall allow unambiguous identification of the audio essence.

6 Bed Metadata

6.1 General

The following subsections specify metadata requirements associated with Beds.

6.2 Bed Identifier

This Identifier uniquely identifies the Bed. No two Beds shall have the same Bed Identifier at any instant in time. This allows referencing multiple Beds if desired.

6.3 Bed Channel List

This metadata item lists the Audio Channels in the Bed.

6.4 Remap coefficients.

Remap coefficients specify how to map the original channels to a different target configuration. This is a set of values indicating how much gain is to be applied to each Audio Channel of a Bed to generate each output Audio Channel of the Bed for the target Soundfield Configuration.

If remap coefficients are provided, additional metadata shall identify the conditions (Target Environments) under which they should be used.

6.5 Conditional Bed

A mixer may want to include one or more alternative Beds for different target Soundfield Configurations. The metadata shall support identifying alternative Beds and the conditions (Target Environments) under which they should be used.

At a minimum, the following target Soundfield Configurations, defined in SMPTE ST 428-12, shall be supported:

1. 5.1: (L, R, C, LFE, Ls, Rs)
2. 7.1DS: (L, R, C, LFE, Lss, Rss, Lrs, Rrs)

7 Object Metadata

7.1 Introduction

An Audio Object is a set of audio samples and associated metadata intended for reproduction according to the position in space and other properties as indicated by the metadata. The position may or may not be associated with a single Loudspeaker. The following subsections specify metadata requirements associated with Audio Objects. Note that all metadata are considered to be dynamic (time-varying) unless explicitly stated differently

7.2 Object Identifier

This metadata item provides a unique identity for an object. No two objects shall have the same Object Identifier at any instant in time. An Object Identifier is static for the duration of an object. The cardinality of the identifier space shall be large enough to support the highest number of simultaneous Audio Objects that exist at any point in time during the presentation.

7.3 Waveform Reference

This metadata item references audio essence associated with the Audio Object. The reference shall allow unambiguous identification of audio essence. The Waveform Reference may include a (non-zero) sample offset.

7.4 Object Position

This metadata item locates the Audio Object in a three-dimensional space. The coordinate system and frame of reference to be used in this standard is defined in Section 9.

7.5 Object Spread

This metadata element describes the Audio Object's size and shape in a three-dimensional space. The coordinate system and frame of reference to be used in this standard is defined in Section 9.

7.6 Object Gain

This metadata item specifies the gain applied to audio essence associated with the Audio Object and shall allow a gain of 0 dB.

7.7 Object Lifetime

An Audio Object lifetime may be explicitly included. The metadata item specifies the intervals (start and duration for a single interval) in time when an Audio Object is active (or present), i.e. the intervals in time where Audio Object metadata may impact Audio Object rendering.

7.8 Object Audio Description

This metadata describes a characteristic of the Audio Object that may be used to target the Audio Object for processing. Supported types shall include: dialog, music, and effects.

7.9 Decorrelation

The perceived sound image when reproducing an audio object across two or more Loudspeakers in an immersive sound system can be localized or diffuse depending on whether the source signals representing the audio object are correlated or decorrelated, respectively. Reproducing a sound with multiple correlated signals yields an easily locatable sound image. Reproducing a sound with multiple related but uncorrelated sources yields a broader, more diffuse sound image. This can be called a decorrelation effect.

The decorrelation metadata item refers to processing the source signals used to reproduce an auditory event to alter their relationship while maintaining the original sound for each individual signal. The minimum value indicates that no decorrelation effect is intended and the maximum value indicates that the maximum decorrelation effect is intended.

7.10 Snap Tolerance

This metadata item indicates the degree to which preservation of object timbre has priority over preservation of object position. This property has extreme values indicating ‘preserving object timbre has highest priority’ and ‘preserving object position has highest priority’, respectively.

7.11 Conditional Object

A mixer may want to include one or more alternative Audio Objects for different target Soundfield Configurations. The metadata shall support identifying alternative Audio Objects and the conditions (Target Environments) under which they shall be used. Note: A conditional Audio Object can replace another Audio Object.

7.12 Zone control

7.12.1 General

This metadata item, when present, indicates the degree to which specified set of Loudspeakers (Zones, as defined in Table 1) are to be excluded from rendering. The actual mapping of Zones to Loudspeakers will be defined, for a cinema, at the time of Renderer configuration. Zones should be configured as a non-overlapping partition of the set of all available Loudspeakers.

7.12.2 Zone Gain

This metadata item represents the degree to which a Zone is included in sound reproduction. This property shall support a range of values. The range shall include extreme values indicating ‘fully enabled’ and ‘fully disabled’. An Audio Object may have a separate gain value for each basic Zone as defined in Table 1. See SMPTE ST 428-12:2013 and SMPTE ST 2098-5 for definitions of speaker locations and layers.

Table 1 - Zones

Number	Description
1	Base layer screen Loudspeakers left of center
2	Base layer center screen Loudspeakers
3	Base layer screen Loudspeakers right of center
4	Height layer screen Loudspeakers left of center
5	Height layer center screen Loudspeakers
6	Height layer screen Loudspeakers right of center
7	Base layer rear wall Loudspeakers left of center
8	Base layer center rear wall Loudspeakers
9	Base layer rear wall Loudspeakers right of center
10	Height layer rear wall Loudspeakers left of center
11	Height layer center rear wall Loudspeakers
12	Height layer rear wall Loudspeakers right of center
13	Base layer left wall Loudspeakers
14	Height layer left wall Loudspeakers
15	Base layer right wall Loudspeakers
16	Height layer right wall Loudspeakers
17	Top layer Loudspeakers left of center
18	Top layer center ceiling Loudspeakers
19	Top layer Loudspeakers right of center

8 Structural Metadata

8.1 Group

It is recognized that it may be desirable to identify certain Audio Objects or Audio Channels as belonging to a Group for the purposes of common processing or action. Immersive audio metadata may support the function of allowing Audio Objects or Audio Channels to be associated with a Group for this purpose. The method by which this is done is not defined.

9 Coordinate System and Frame of Reference

9.1 Coordinate System

Audio Object positional metadata shall indicate the placement of an Audio Object using a Cartesian coordinate system. This system uses three orthogonal axes, (x, y, z), to locate a point in space with respect to a chosen origin. The x coordinate shall represent the left-right dimension, the y coordinate shall represent the front-back dimension, and z coordinate shall represent the down-up dimension as shown in figure 1.

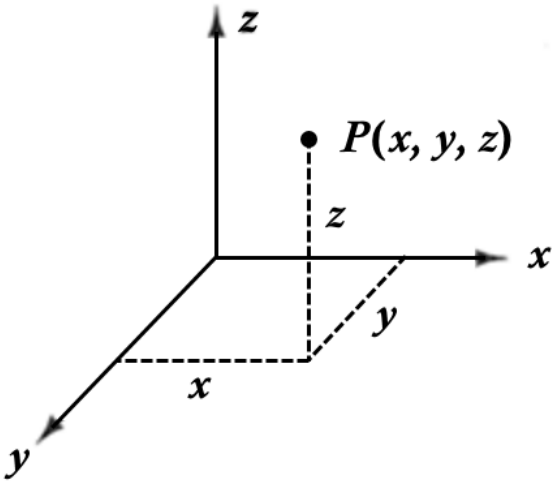

Figure 1. Cartesian Coordinate System

9.2 Frame of reference and location coding

The Cartesian coordinate values used for Audio Object position shall be normalized relative to reference points of a cube, which represents an idealized cinema model. The front plane is the location of the screen; “left” is relative to an observer in the cube, facing the front. The metadata may support Audio Object locations inside, on, and outside the cube. At a minimum, location metadata shall support locations on and inside the cube from the Z axis midpoint to the top of the cube. The reference points are defined below.

For the x-axis, the reference points shall be the left and right cube faces. These reference points shall have fixed, defined values.

For the y-axis, the reference points shall be the front and back cube faces. These reference points shall have fixed, defined values.

For the z-axis, the reference points shall be the bottom, midpoint, and top of the cube. These reference points shall have fixed, defined values.

Locations within the cube shall be represented using linear interpolation between the adjacent reference points on each axis. Locations outside the cube, if supported, shall be represented based on a linear extrapolation from the reference points on the cube.

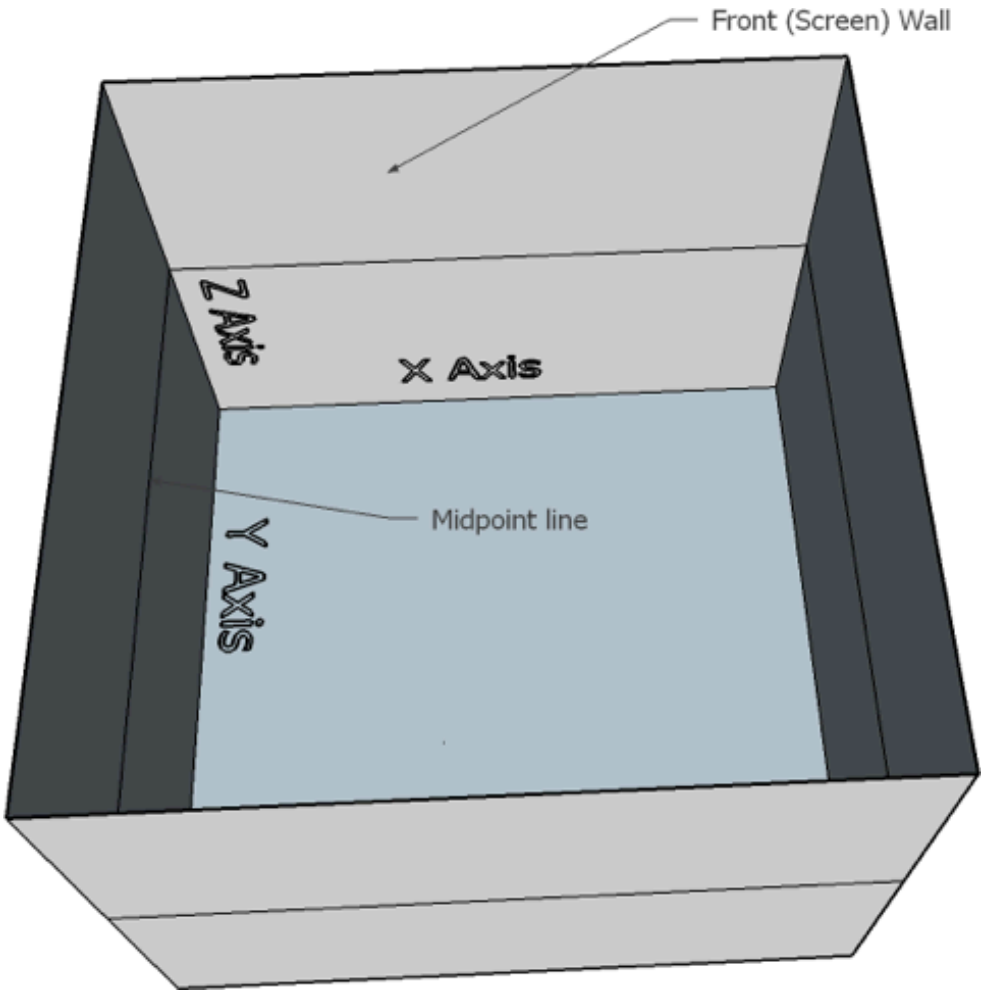

Figure 2 - Cube Showing Some Reference Points

9.2.1 Example

The actual cube reference values are to be determined as part of the bitstream definition and are outside the scope of this document. The following is one possible definition.

The origin, [X,Y,Z] = [0,0,0], is set to be the bottom left front corner of the cube. The defined values are as follows:

X axis: left face value = 0; right face value = 1

Y axis: front face value = 0; rear face value = 1

Z axis: bottom value = 0; top value = 1; midpoint value = 0.5

A coordinate of [0.5, 0, 0.5] will be located on the center of the front face.

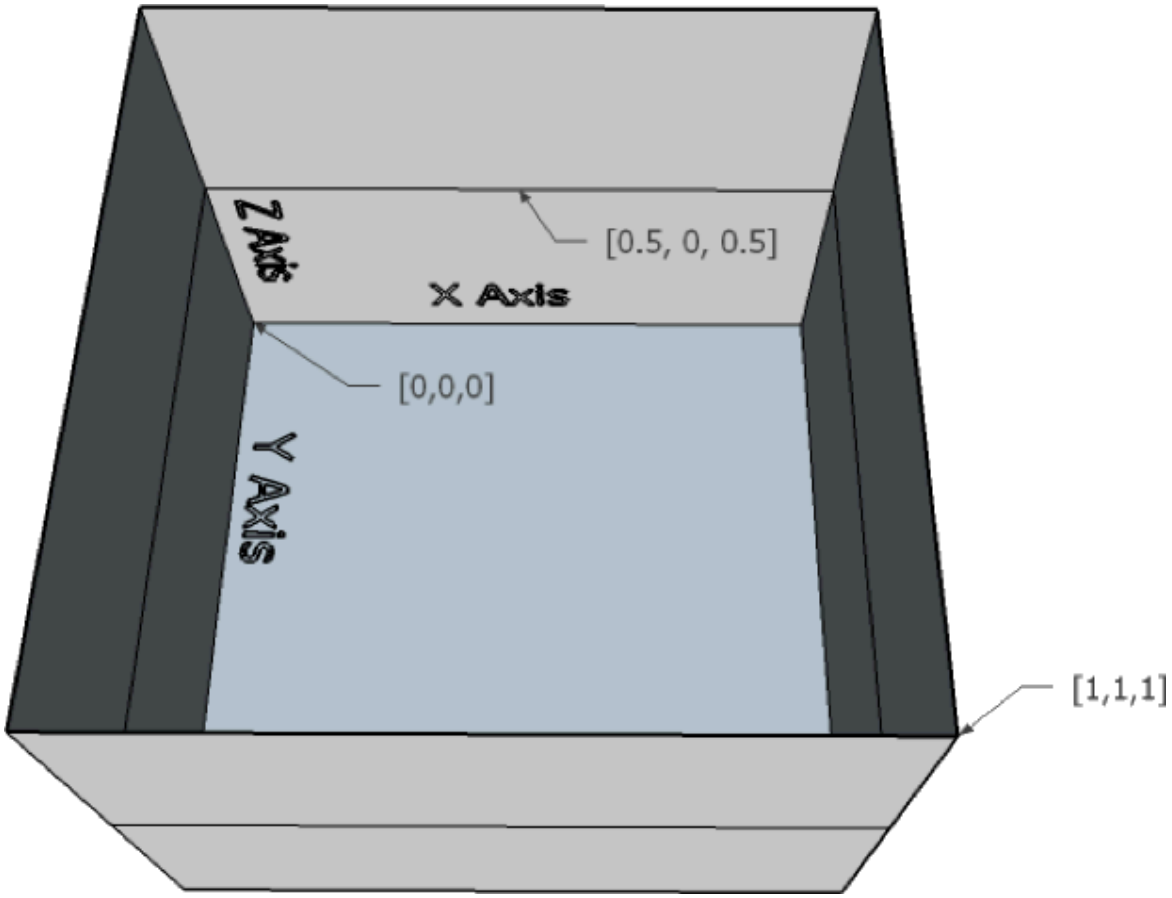

Figure 3 - Cube with Example Coordinates

9.3 Mapping the cube to a cinema

Mapping of Object Positions within the cube to cinema (or mixing stage) Loudspeakers is the function of the Renderer and is out of scope of this document. However, the reference points on the cube shall have defined meaning regardless of the shape of the room into which the locations are mapped.

The front face of the cube shall map to the nominal front wall of the cinema. Furthermore, the frontmost Loudspeaker-mounting surface shall be considered the nominal front wall of the cinema.

The left face of the cube shall map to the nominal left wall of the cinema. Furthermore, the leftmost Loudspeaker-mounting surface shall be considered the nominal left wall of the cinema.

The right face of the cube shall map to the nominal right wall of the cinema. Furthermore, the rightmost Loudspeaker-mounting surface shall be considered the nominal right wall of the cinema.

The back face of the cube shall map to the nominal rear wall of the cinema. Furthermore, the rearmost Loudspeaker-mounting surface shall be considered the nominal rear wall of the cinema.

The mid-height plane of the cube shall map to the height of a legacy 2-dimensional Loudspeaker system (e.g. 5.1 or 7.1).

The top of the cube shall map to the ceiling of the cinema. Furthermore, the overhead Loudspeaker-mounting surface shall be considered the nominal ceiling of the cinema.

The bottom of the cube shall map to the nominal floor of the cinema.

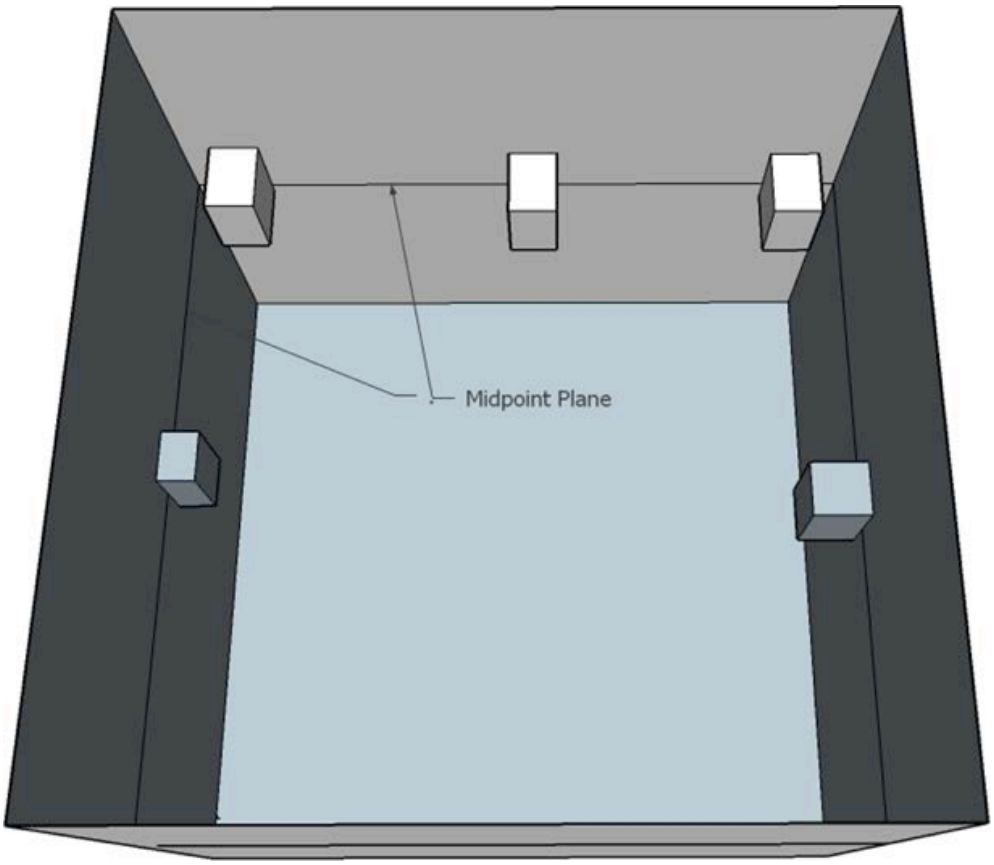

Figure 4 - Mapping of Midpoint Plane To Base Layer Loudspeaker Height

9.4 Example - Mapping object position to a cinema Loudspeaker array (informative)

As mentioned in section 9.3, mapping of object positions within a cinema is the function of the renderer and is out of scope of this document. The following example is provided to assist the reader in understanding some of the issues involved in mapping object location into an example cinema. This example should help illustrate the point that, by using reference points that map to any room, the intent is not to preserve object positions relative to an observer, but relative to the shape and size of the playback environment.

Figure 5 shows two diagrams. The one on the left shows a possible panner interface, the one on the right shows the simplified overhead view of a cinema. In the panner diagram, the mixer has placed a sound source (green dot) 1/3 of the way from front to back, and 1/4 of the way from left to right. The light blue line represents the room perimeter of a nominal, square room.

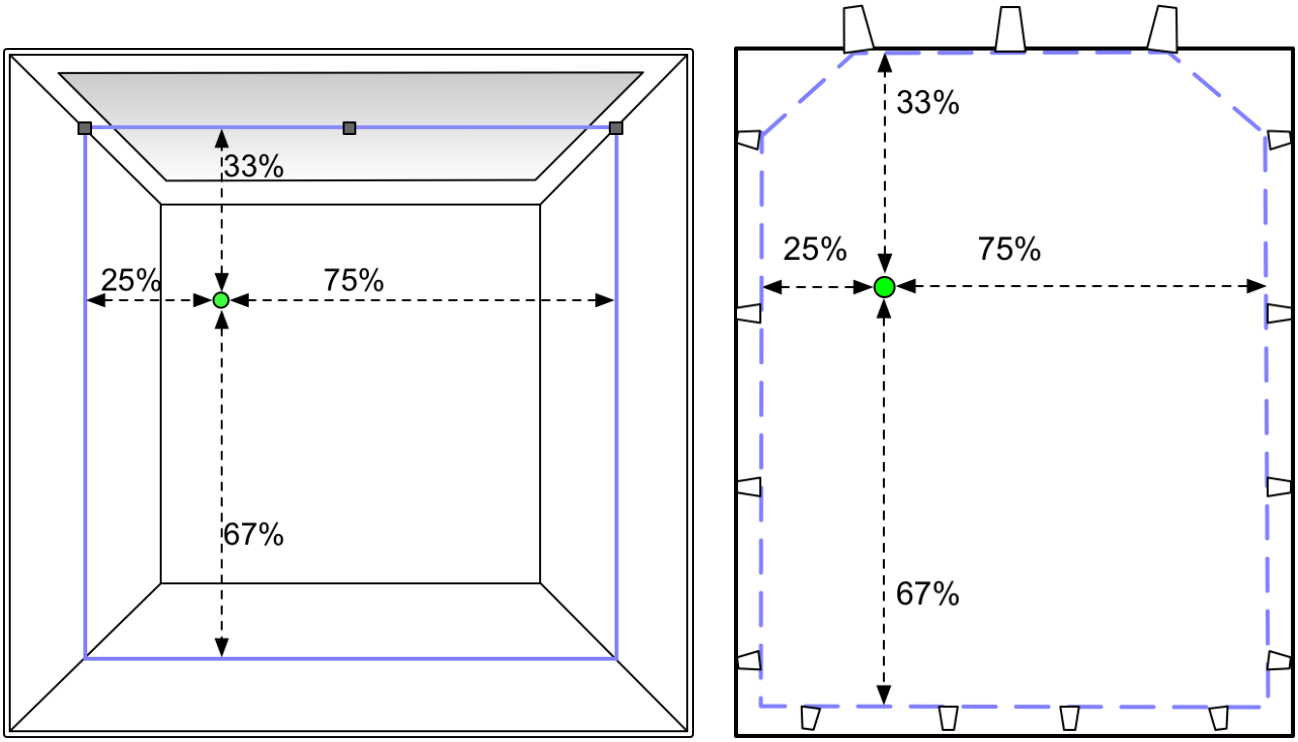

Figure 5 - Mapping ideal cube to example cinema

The cinema diagram shows how the object position could be mapped to a cinema that is not a square, but is longer than wide, and has the screen left and right Loudspeakers mounted inside the left and right walls. The dotted-blue line shows the perimeter of the room as now defined by the Loudspeaker mounting. This will vary from cinema to cinema. Note that the position relative to the reference points (walls) remains the same. The sound source is 1/3 of the way from front to back, and 1/4 of the way from left to right.

Bibliography (Informative)

Report of the Study Group on Immersive Audio Systems: Cinema B-Chain and Distribution, SMPTE Report, March 2014.