Recent Advancements in Audio – How a Paradigm Shift in Audio Spatial Representation & Delivery Will Change the Future of Consumer Audio Experiences

Jeffrey C. Riedmiller and Nicolas Tsingos
Dolby Laboratories Inc.

Abstract

Recent advancements in spatial audio processing, distribution and delivery are now capable of bringing more lifelike, scalable and interactive audio experiences to consumers across a wide range of devices and applications than ever before. This paper provides the reader with a high-level tutorial on the latest research and development supporting the creation, distribution and delivery of immersive and personalized audio experiences to consumers.

BACKGROUND

Over the past 70+ years, broadcast television sound has evolved from providing a single *complete* (monophonic) audio program to having the capability of supporting a combination of 5.1-channel and stereo sound tracks, including multiprogram support for delivering alternate languages and/or descriptive audio. While the digital transition enabled efficient delivery of multiple audio programs to consumers, television sound continues to be created, distributed and delivered as one or more *complete* audio programs just as it has been for decades. The broadcast standards community (including the bitstream syntax utilized for consumer delivery) typically refers to these audio service types as *Complete Main* services. *Complete Main* services contain a complete audio program including dialogue, music and effect elements *mixed* together in either 5.1- or 2.0-channel format *prior* to being broadcast to consumers. Moreover, additional languages and/or descriptive programming are also *mixed* with dialogue, music and effects elements *prior* to being broadcast as *Complete Main* service types as well. Today, it is quite common for television services to include two audio programs (i.e. two elementary compressed bitstreams multiplexed with video and other service information). The main (primary) audio program in the native language is typically offered in 5.1-channel format while the secondary language is typically provided in stereo. Descriptive audio programming, when available, typically replaces the secondary language with native language descriptions pre-mixed over a stereo version of the native language soundtrack – restricting the availability of secondary language programming.

The primary and secondary audio program examples above provide viewers with two audio ***presentations*** in two-dimensional (5.1) and one-dimensional (stereo) channel-based formats respectively; where the primary program is offered in 5.1 (two-dimensional) and the secondary program is offered in 2.0/stereo (one-dimensional). The term ***presentation(s)*** is a key term associated with next-generation consumer audio experiences that aim to offer a greater degree of ***personalization*** under viewer control along with providing non-native speaking, hearing-impaired and visually-impaired audiences with the highest quality experience (only available with the primary audio program today).

To address the currently limited set of audio presentations (and related experiences), recent advancements in the development and

deployment of spatial audio [1][14][15] systems are poised to make the personalization of audio a practical reality for content creators, programmers, MVPDs (Multichannel Video Programming Distributors) and consumers alike. Additionally, these advancements also provide a substantial step forward in terms of ***immersiveness*** by enabling sound to be represented and delivered to consumers in three-dimensions. Immersive sound research and formats are fundamentally focused on providing a more accurate representation of the spatial attributes of an audio program (in a higher-dimensional space) while providing listener experiences that engage our sensory system in a more natural way – allowing sensory systems to process information more similarly to how listeners experience the natural world by providing more realistic and natural auditory cues to the listener including advanced binaural for a headphone listener.

IMMERSIVE SOUND – OBJECT-BASED AUDIO

The object-based audio paradigm provides the foundation for immersive sound and consists of a set of monophonic audio elements (tracks) and tightly coupled metadata representing the object's position in three-dimensional space. The positional metadata is generated with a great enough frequency to enable dynamic object movement and precise localization during playback. At playback time, an audio rendering engine utilizes this metadata to map each of the audio objects to a single or a combination of speaker outputs to achieve the desired spatial effect. Importantly, each object's position is represented as a coordinate in three-dimensional space that can also leverage overhead and/or upward-firing speaker configurations in addition to advanced 3D headphone rendering. For instance, ITU-R BS.2051 [10] defines up to 22 primary channels that can be used for playback. In the home, we expect 5.1 + 2 overhead speakers or 7.1 + 4 overhead speakers to be the most widely adopted 3D loudspeaker configurations.

Today's channel-based audio paradigm along with the emerging object-based audio paradigm are alternative methods for describing sound events. Channel-based audio expresses the audio scene in terms of loudspeakers utilizing pre-determined (nominal) positions and characteristics. Object-based audio describes sound events in a way that is independent of loudspeaker configurations/layouts and positions. The spatial accuracy, homogeneity and resolution possible with object-based audio enable new possibilities for the sound designer, mixer and director.

In traditional channel-based audio mixing, sound elements are mixed together using fixed speaker positions and angles assumed to be identical in every playback environment. For instance, 5.1-channel audio productions are created utilizing the left, right, center, left surround, and right surround speaker layout conforming to ITU-R BS.775. However, 5.1-channel playback in consumer environments typically deviates in terms of speaker distance and angles relative to the listener's position from the ITU-R BS.775 recommendation. This deviation has a substantial impact on the consumer's ability to experience the soundtrack in a manner that was intended (and heard) by the program producer during production. On the other hand, object-based audio mixing enables ***individual*** or ***groups*** of sound elements to be described in a three-dimensional space (rather than being forever assigned to a specific channel/speaker location) and delivered to the playback device where they are ***rendered*** based on the active speaker layout in use, including the number and position of speakers in 1-, 2- or 3-dimensions. The benefit of this approach is that by decoupling audio mixing and spatial representation from a single fixed channel-based layout, the audio soundtrack/experience

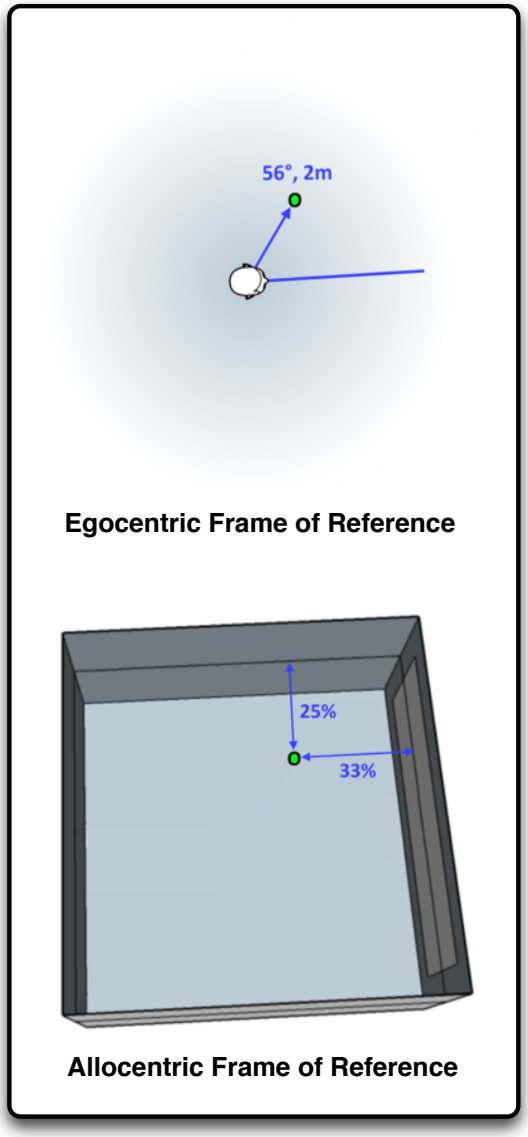
can scale optimally across any speaker layout in use today or in the future. In other words, the creative process utilizing objects mixes for the 'space' rather than to a fixed speaker location – again, which is assumed to be identical in every playback environment.

Coordinate System & Frame Of Reference

When designing a system to enable individual audio elements (objects) to be represented in 3-dimensional space, a frame of reference is required. Typically a system designer will consider the distinction between an *egocentric* (observer) or an *allocentric* (environmental) reference as a fundamental consideration.

An egocentric frame of reference represents (or encodes) an audio object's location relative to the position of the observer (or 'self'). An egocentric reference is commonly used for the study and description of perception.

An allocentric frame of reference represents (or encodes) an audio object's location using a reference location and direction relative to other objects in the environment. An allocentric reference is better suited for a scene description that is independent of a single observer's position and when the relationship between elements in the environment is of interest.



The rationale for choosing a frame of reference should consider the following:

- Provide optimum capture of artistic intent
- Provide optimum translation across a variety of listening environments.
- Be consistent with the production tools utilized to capture artistic intent.
- Provide consistent and predictable behavior across a wide listening area.

In general, audio mixers naturally approach the creative process in allocentric terms that is further reinforced by the fact that

audio panning tools having been implemented with an allocentric frame of reference for decades (i.e. the screen, room left, room back, etc.). As an example, the mixer's thought process for rendering audio objects follows "this sound should be on-screen", "this sound should be off-screen and a ¼ of the way from the left to the right wall", etc. Thus, audio object movement in space is defined in relation to the playback environment, e.g. a fly-over from the center of the screen, up across the ceiling and ending at the center of the back wall. Using an allocentric frame of reference, all of these relationships are preserved.

In contrast, utilizing an egocentric frame of reference can, for example, yield results where an object placed on the side wall of a large rectangular mixing stage ending up on the rear wall of a more square space during playback. Moreover, with three-dimensional egocentric audio frameworks (which includes distance) an object on the rear wall of a small mixing stage could end up well within the audience area of a large playback space such as an auditorium.

Artistically, the location of an audio object is most important when associated with an on-screen visual object. Using an allocentric reference and representation for every listening position, and for any screen size, individual sound objects can be described at the same relative position on the screen and optimally reproduced across a wide range of room sizes and shapes found from cinema to home environments.

Considering the rationale behind the use of an allocentric frame of reference, an audio object's position is preferably defined as a unit room where each object(s) 3-dimensional coordinates (x,y,z) in the range of [-1,1] x [-1,1] x [-1,1], correspond to the traditional panpot controls found in mixing consoles - left/right, front/back and by extension to 3D bottom/top. This is commonly referred to as a cartesian room normalized coordinate system, where X= -1 is full left, X=1 is full right, Y= -1 is front, Y=1 is back, Z=0 is traditional surround plane, Z=1 is overhead plane. This can be extended to Z= -1 to include a floor plane (below the listener). This defines and describes a normalized room hemicube/cube as in Figure 2.

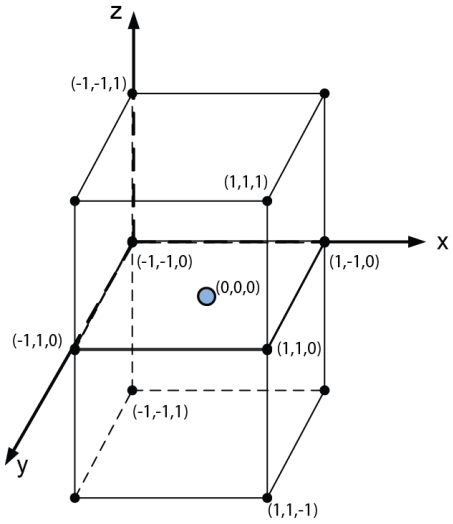


Figure 2 : Object Position Coordinate System

Modern surround-sound cinema and object-based consumer delivery systems use an allocentric frame of reference. ITU-R BS.2051 [10] provides mapping coordinates for predefined channels and corresponding loudspeaker positions associated with advanced sound reproduction systems.

Object-based Audio Rendering

An audio object rendering engine is a mandatory requirement for playback products and applications that support object-based audio for immersive and personalized audio experiences. An audio renderer converts a set of audio object signals with associated metadata to a different configuration of audio signals – e.g. speaker feeds, based on the metadata, **AND** a set of control inputs derived from the rendering environment and/or user preference.

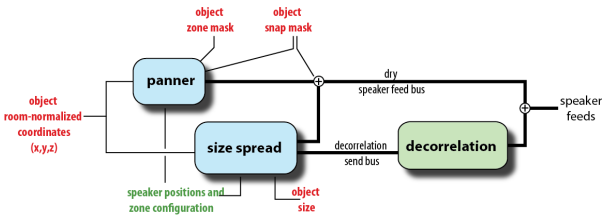


Figure 3 – Object-based Audio Renderer

At the core of rendering is a panning algorithm (see Figure 3). Most panning algorithms currently used in object-based audio production attempt to recreate the original auditory cues during playback via amplitude panning techniques [5][6][7][8] where, a gain vector [**Gi**] is computed and assigned to the source signal for each of the *n* loudspeakers in use The object signal s(t) is therefore reproduced by each loudspeaker as Gi(x,y,z)*s(t) in order to recreate suitable localization cues as indicated by the object <x,y,z> coordinates.

The design of panning algorithms ultimately has to balance tradeoffs among timbral fidelity, spatial accuracy, smoothness and sensitivity to listener placement in the listening environment, all of which can affect how an object at a given position in space will be perceived by listeners [8][9]. For instance, Figure 4 illustrates how different speakers maybe utilized among various rendering (panning) algorithms to place an object's perceived position in the playback environment.

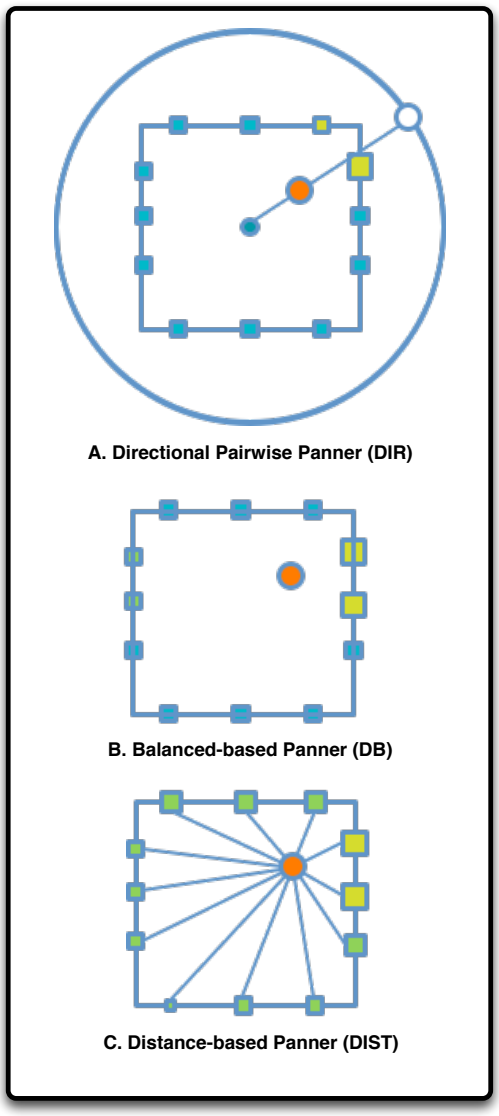


Figure 4 – Common Panning Algorithms

Directional pairwise panning (DIR) (see Figure 4 (a)) is a commonly used strategy that solely relies on the directional vector from a reference position (generally the sweet-spot or center of the room) to the desired object's position. The pair of speakers 'bracketing' the relevant directional vector is used to place (render) that object's position in space during playback. A well-documented extension of directional pairwise panning to support 3D loudspeaker layouts is vector-based amplitude panning (VBAP) [7] which uses triplets of speakers. As this approach only utilizes the direction of the source relative to a reference

position, it cannot differentiate between object sources at different positions along the same direction vector. It can also introduce instabilities as objects are panned near the center of the room. Moreover, some 3D implementations may constrain the rendered objects to the surface of a unit sphere and thus would not necessarily allow for an object to cross inside the room without going 'up and over'. As a side note, directional panning solutions can also create sharp speaker transitions as objects approach the center of the room. Where, a small movement an object's position would not always translate into a small variation in loudspeaker gains [Gi] for objects moving (panning) *through* the room/space.

The 'dual-balance' panning algorithm is the most common approach used in 5.1/7.1-channel surround productions today (Figure 4 (b)). This approach utilizes left/right and front/back panpot controls widely used for surround panning. As a result, dual-balance panning generally operates on the set of four speakers bracketing the desired 2D object position.

Extending to three-dimensions (e.g., when utilizing a vertical layer of speakers above the listener) yields a "triple-balance" panner. It generates three sets of one-dimensional gains corresponding to left/right, front/back and top/bottom balance values. These values can then be multiplied to obtain the final loudspeaker gains:

$$Gi(x,y,z) = Gx\_i(x) \times Gy\_i(y) \times Gz\_i(z).$$

This approach is fully continuous for objects panned across the room in either 2D or 3D and makes it easier to precisely control how and when speakers on the base or elevation layer are to be used.

In contrast to the directional and balance-based approaches, distance-based panning (DIST) [5] (Figure 4(c)) uses the relative distance from the desired 2D or 3D object location to each speaker in use to determine the panning gains. As a result, this approach generally utilizes all of the available speakers in use rather than a limited subset which leads to smoother objects pans but with the tradeoff of being prone to timbral artefacts.

It should be noted that both 'dual balance' and 'distance-based' panning support smooth object pans in the sense that a small variation in an object's position will translate to a small change in loudspeaker gains.

Object-based Audio Rendering Requirements

Any object rendering algorithm used in professional or consumer playback applications must address (at a minimum) the following high-level functional requirements:

1. Support smooth panning for object rendered inside and through the room – including the two-dimensional surround plane and the three-dimensional cube/hemicube described above.
2. Support for rendering utilizing a traditional two-dimensional surround speaker layout even when overhead speakers are present. e.g. objects with z = 0 will render across the 2D speaker layout while objects with z = 1 will only render via overhead/upward firing speakers.
3. Implement an approach that minimizes the number of speakers used to reproduce an object – to preserve timbral fidelity.
4. Support additional creative control (via metadata) including the ability to dynamically mask out some of the speakers (or zones of speakers) when an object is being rendered across specific loudspeaker configurations
5. Be power preserving across speaker layouts: $\sum i\ Gi^2 = 1$.

Supporting the requirements above, the authors recommend a triple-balance implementation for deriving loudspeaker gains from object coordinates. The triple-balance approach expresses speaker gains **Gi** as the separable product of three panpot controls: left/right, front/back, top/bottom and is a direct extension of the commonly available panning tools on mixing consoles. The following background and supporting equations provide a representative example of how a simple panning algorithm would be implemented using a simple sine/cosine law. Other panning laws are also possible.

An indicative example for computing a one-dimensional (stereo) set of rendering gains utilizing an audio object's x-coordinate in [-1,1] could be derived as follows:

$$G\_left = \cos((x+1)/2.0*\pi/2)),$$
$$G\_right = \sin((x+1)/2.0*\pi/2))$$

An indicative 3/2/0 example (L/C/R/Ls/Rs) would be (2D rendering using <x,y> coordinates in [-1,1] x [-1,1]) :

A) Set all gains to 0.0

B) Compute left/right panpot for front and back speakers as in the previous stereo example:

```
G_ls = cos((x+1)/2.0*π/2)
G_rs = sin((x+1)/2.0*π/2)

if (x<=0.0f)
{
 G_l = cos(-x*π/2);
 G_c = sin(-x*π/2);
}
else
{
 G_c = cos(x*π/2);
 G_r = sin(x*π/2);
}
```

C) Combine with front/back panpot:

```
float c = cos((y+1)/2.0*π/2);
float s = sin((y+1)/2.0*π/2);

G_l *=c;  G_r *=c; G_c *=c;
G_ls *=s;  G_rs *=s;
```

D) Normalize power to 1.0 by dividing all gains Gi by $sqrt(\Sigma i\ Gi^2)$

Following the same principle, these examples can be easily extended using a third dimension for elevation (height).

Artistic Control Metadata

While the use of a consistent core rendering algorithm is desirable, it cannot be assumed that a given solution will always deliver consistent and aesthetically pleasing results across different playback environments. For instance, today the production community commonly remix the same soundtrack for different channel-based formats in use worldwide, such as 7.1/5.1 or stereo, to achieve their desired artistic goals for each format. With several hundred audio tracks competing for audibility, maintaining the discreteness of the mix and finding a place for all the key elements is a challenge that all theatrical/TV mixers face. Achieving success often requires mixing rules that are deliberately inconsistent with a physical model or a direct re-rendering across different speaker configurations.

To solve this problem, we introduce additional metadata used to dynamically reconfigure the object renderer [11] (Figure 3). For instance, the metadata can be used to "mask out" certain speaker zones during playback or use only a single loudspeaker to render an object for improved timbral fidelity. Alternatively, metadata can also be provided to render wide objects, covering a large spatial extent. Implementation of wide objects

requires the use of decorrelation processing in order to recreate the proper cues for perceptual size.

PERSONALIZED SOUND & ACCESSIBILITY

Audio personalization is a new concept that aims to enable viewers to choose and, when appropriate based on the intent of the content creator, tailor the experience to their unique needs or preferences. Personalization can include selection of languages, alternative commentary, visual descriptions along with the ability to adjust the level (and in some cases the spatial) relationship between the dialogue-based and music/effects audio object elements. Extending the object-based audio paradigm to support the professional interchange and consumer delivery of both individual audio objects and groups of audio objects (with accompanying metadata) will provide the building blocks for enabling the personalization of audio programming [12]. It will also ensure that the entire audience, including native and non-native speaking, hearing and visually impaired, are provided with a premium (including an immersive) experience typically only available with the primary language audio program today. As a simple example, the primary object-based audio building blocks (including metadata) for a program can be defined as represented in Figure 5.

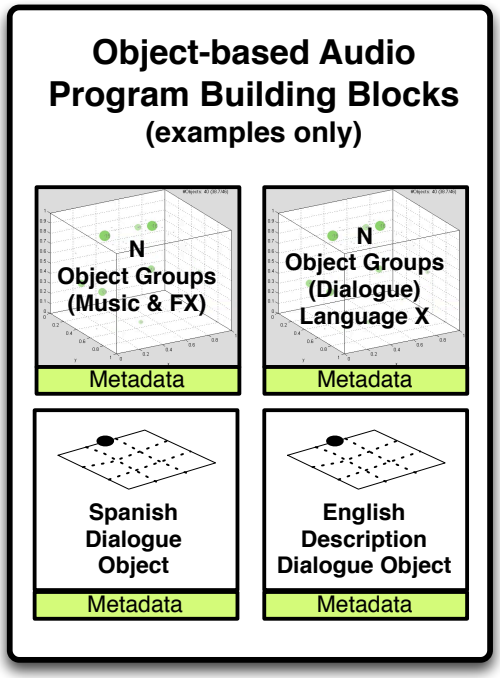


Figure 5 – Objects & Object Groups

The Music & FX object group (as the name implies) contains a set of spatial audio objects and metadata representing only the music and sound effects for the program. As represented in Figure 5, this object group when being simultaneously rendered with the English Dialogue object group would represent a single immersive *presentation* for the consumer. While the combination of the Music & FX object group and the Spanish dialogue (represented here as a single object rather than an object group with spatial attributes including spatialized effects - reverb, etc.) represents a second immersive *presentation.* A third *presentation* is further represented by the simultaneous rendering of the Music & FX group, English object group and the English description object. Again, these elements can be rendered based on the preferences of the consumer/viewer.

This approach provides more efficient delivery and higher quality experiences for non-native speaking and visually- and hearing-impaired audiences by supporting a fundamentally different approach to

programming distribution. Instead of distributing complete mixes of each type (for example, hearing- or visually-impaired), this approach distributes the program building blocks as discrete objects or object groups each containing music and effects, main dialogue, alternate dialogue, dialogue descriptions, etc. These program building blocks are then dynamically rendered during playback, possibly with consumer adjustment (personalization – including adjustment of the relative level between dialogue objects and the Music & FX object group – this address the needs of the hearing impaired audience). This approach also improves efficiency by reducing the overall datarate required for delivering the complete set of possible presentations.

A FLEXIBLE CONSUMER DELIVERY FORMAT FOR IMMERSIVE & PERSONALIZED AUDIO

Enabling efficient and flexible delivery of immersive and personalized programming to consumers also requires a new approach for consumer delivery. The AC-4 format standardized in ETSI TS 103 190 [13] provides a state-of-the-art and efficient mechanism for delivering both personalized and immersive audio experiences in object-based or traditional channel-based paradigms.

The AC-4 bitstream consists of a sequence of synchronization frames as shown in Figure 6. Each synchronization frame begins with a synchronization word and ends with a CRC word. The synchronization word enables a decoder to easily identify frame boundaries and begin the decode process while the CRC word allows a decoder or other system component to detect the occurrence of bitstream errors and invoke concealment as necessary. The portion of the AC-4 synchronization frame containing all of the audio essence and metadata is referred to as the raw AC-4 frame.

The high-level AC-4 bitstream structure shown in Figure 6 supports the carriage of audio and metadata substreams described by the Table-of-Contents (TOC) bitstream element. The arrows indicate the order of arrival of the associated information in the AC-4 bitstream.

AC-4 Table of Contents Element

The TOC bitstream element contains all of the information necessary to describe and enable immersive and personalized experiences including support for future extensibility. The *Presentation Info* element contains information describing one or more substreams to be decoded and presented simultaneously. Substreams can be thought of as single decodable units representing a specific channel, group of channels, a specific object or a group of objects. An important concept to note here is that a single substream may be shared across multiple presentations.

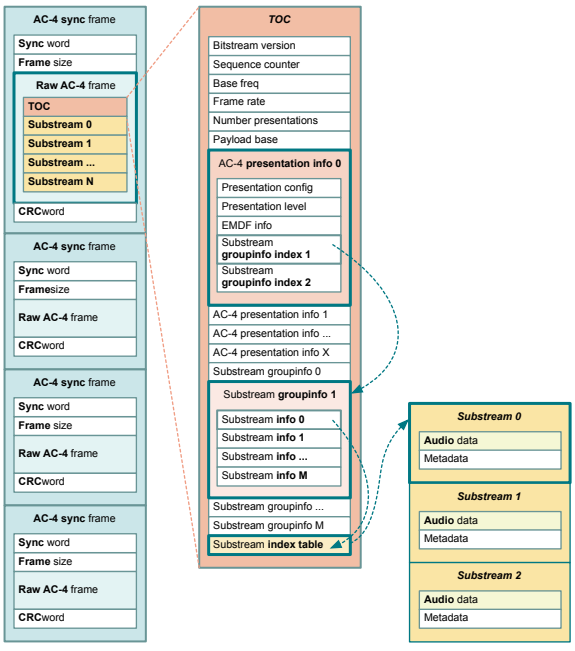


Figure 6: AC-4 Bitstream Syntax – TOC

AC-4 Presentation Information

The AC-4 Presentation Information element (AC-4 presentation info) contains the following sub-elements:

Presentation Config - Defines the configuration of the *Presentation* made up of a group of substreams (such as Music-and-Effects plus Dialogue, or Complete Main plus Associated Audio).

Presentation level - Defines the minimum level of the decoder that is able to decode the *Presentation*. For more information on the decoder levels refer to the AC-4 standard. [13]

EMDF info - Defines the properties of an Extensible Metadata Delivery Format (EMDF) substream associated with the *Presentation*.

Substream group info index – An integer that points to specific Substream Group Info fields dependent on the configuration of the *Presentation Config*.

AC-4 Substream Group Info

The *substream group info* element contains one or more substream info elements containing substream numbers that range from 0 to M. Each substream number points to an index in the substream index table. In the case of a single substream, as needed for a Complete Main channel-based configuration, only the first field in the list of substream info elements is used and the number contained in the substream info points to the respective substream via the substream index table. In the case of coded objects or object groups – for example, eight – the substream info list could be filled with eight substream info elements that point to the eight different objects represented by eight substreams. The substream index table is a table which maps substream numbers to the entry point for a specific substream in the AC-4 frame. All of the payload information such as audio essence and metadata is carried in substreams. When *Substreams* and *Presentations* are being described, each substream in an AC-4 bitstream is given a substream ID corresponding to an integer starting from 0. The first substream is referred to as S0, the second as S1, and so on. The first Presentation info element is referred to as P0, the second as P1, and so on.

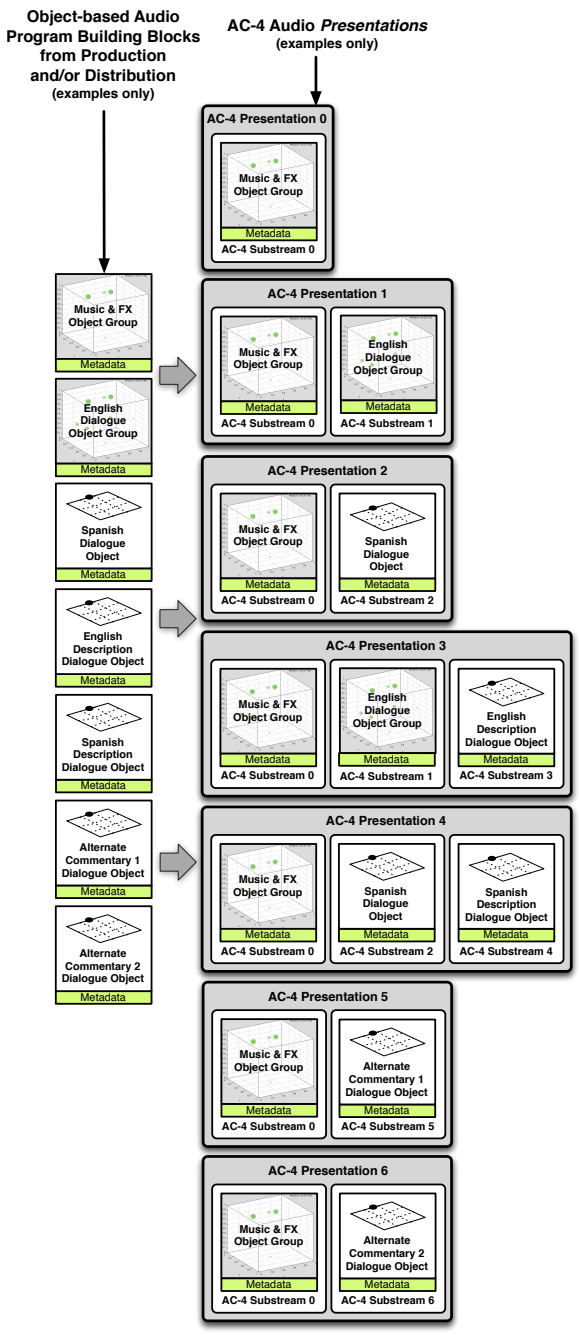


Figure 7 – Object-based building blocks mapped to AC-4 substreams & presentations (example only)

The object and object group building blocks - created during production - shown in Figure 7 are each mapped and carried in the AC-4 bitstream as an individual substream. These substreams are logically grouped into *Presentations* prior to being multiplexed for delivery. The *Presentations* expressed in Figure 7 represent a total of 7 unique *Presentations* carried within a single AC-4 bitstream. This approach provides more efficient delivery of alternate language and descriptive services (including alternate commentary) ALL in an immersive (3D) format. Note the use of the same Music & FX substream is shared across multiple presentations for optimum efficiency.

The example stream in Figure 7 provides the end user (consumer) with 7 unique experiences including:

1. Immersive (3D) Music & FX
2. Immersive (3D) English
3. Immersive (3D) Spanish
4. Immersive (3D) English with English Description
5. Immersive (3D) Spanish with Spanish Description
6. Immersive (3D) with Alternate Commentary 1
7. Immersive (3D) with Alternate Commentary 2

A typical broadcast bitrate for delivering all of the immersive experiences listed above would be <600kbps. In contrast, delivering the same experiences with deployed systems today would require pre-mixing all of the source elements on a presentation-by-presentation basis (prior to encoding) and simulcasting several independent audio elementary streams to viewers. This comes with an excessive cost in terms of total bitrate. (Which in this example is, ~ 2.7 Mbps – assuming 384kbps for each of the 7 presentations available – with the additional restriction of only providing 5.1-channel surround rather than immersive sound in 3D)

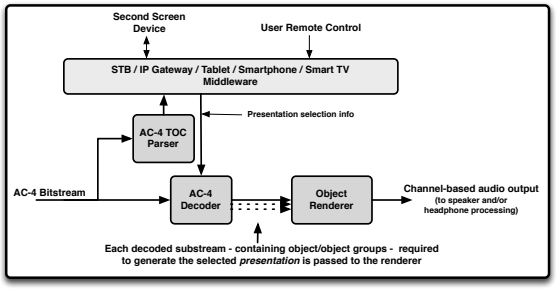


Figure 8 – Representative example of the primary processing blocks in an object-based audio subsystem of a playback device

Playback devices are required to support both traditional channel-based audio (stereo and multichannel) and object-based audio sources. A representative (simple) playback device is pictured in Figure 8. The TOC parser determines the number and type of *Presentations* available in the incoming AC-4 bitstream. It then extracts personalization metadata and decides which audio objects and/or object groups to simultaneously decode based on user selection or will automatically select the default presentation signaled in the bitstream. The playback device only decodes the substream(s) necessary for the selected presentation. The decoder produces a set of audio elements, object-audio metadata, and personalization metadata for use by the object rendering block.

Personalization metadata is further processed by the middleware in the playback device. The middleware presents the different personalization options to the user either via an on-screen user interface or by communicating to a second-screen (e.g. tablet) application. Either a remote control or touch-screen device can be used to customize the audio program. This customization can be as simple as selecting between the different audio presentations or as advanced as selecting the individual audio elements to be presented along with configuring the spatial position and level of each audio object or object group upon rendering. The level and range of audio personalization available is defined by the metadata authored by the content creator and/or distributor.

ADVANCED METADATA

Metadata has long been a fundamental component of the legacy AC-3 system that, for the first time, enabled scalable and consistent playback across the most common devices used at the time. However, as device capabilities, form factors, distribution paths and user applications have expanded over the past 20 years, the need for a more robust and richer set of metadata for this purpose has become increasingly apparent.

To address the device and application needs for today – with the capability to extend in the future - the AC-4 format supports an expanded and enriched set of metadata in many new categories including:

- **Presentation**: logical grouping and labeling of sub-stream/program elements to control playback mixing and rendering
- **Loudness**: ITU-R relative and speech-gated loudness information/control and regulatory information for automated adaptation of processing throughout the delivery and playback chain
- **Advanced dynamic range control**: multichannel and multi-band dynamic range information/control, including downmix and rendering loudness control
- **Dialogue enhancement:** support for both legacy (pre-mixed channels) and next-generation (music-and-effects plus dialogue) program sources
- **Spatial representation**: support for both channel-based and dynamic object-based programming
- **Dynamic rendering control**: speaker and headphone rendering control

- **Program/bitstream identification and synchronization**: for second-stream and second-screen applications
- **Interactive controls:** for personalization including language replacement, visual description and dialogue enhancement

All metadata generation and carriage within the AC-4 system takes advantage of a secure authentication mechanism that ensures robust and reliable delivery throughout the distribution and re-distribution pathways. This mechanism validates both the source and validity of the metadata as trusted. Metadata failing the authentication can be ignored or replaced by an authorized broadcaster.

To address future needs, metadata extensibility is critical to ensure that a low-friction pathway for distributing/delivering new metadata is always available. The Extensible Metadata Delivery Format (EMDF) syntax in AC-4 provides a structured and extensible container for a collision-free and open pathway for additional information (for example, third-party metadata, third-party application data, and so on) to be carried in AC-4 bitstreams and throughout the Dolby Audio System. This makes it easy for content creators and broadcasters to take advantage of this capability.

This following section provides an overview of the metadata parameters (and their application) essential for enabling next-generation immersive and personalized experiences utilizing the AC-4 format. Note: the metadata described in this section is not exhaustive and thus only focuses on the essentials for object-based rendering including personalization. The full set of metadata is described in ETSI TS 103 190 [13].

In general, the primary purpose of the object audio metadata is to:

- Describe the composition of the object-based audio program
- Deliver metadata describing how objects should be rendered
- Describe the properties of each object (for example, position, media classification, etc.)

Within the AC-4 format, a subset of the object audio metadata fields is essential to provide the best audio experience and to ensure that the original artistic intent is preserved. The remaining non-essential metadata fields are used for either an enhanced playback applications or aiding in the transmission and playback of the program content.

Essential Object Metadata for Immersive & Personalized Audio

Metadata critical to ensure proper rendering of objects and provide sufficient artistic control include:

- Object type / assignment
- Timing (timestamp)
- Object position
- Zone / elevation mask
- Object width
- Object snap
- Object divergence
- Number of presentations
- Default presentation
- Number of substreams per presentation
- Substream content type
- Language
- Gain

Object Type / Object Assignment - To properly render a set of objects, both the object type and object assignment of each object in the program must be known. For

spatial objects, two object types defined for current object-based audio production.

- Bed objects - This is an object with positional metadata that does not change over time and is described by a predefined speaker position. The object assignment for bed objects describe the intended playback speaker, for example, Left (L), Right (R), Center (C) … Right Rear Surround (Rrs) … Left Top Middle (Ltm).
- Dynamic objects - A dynamic object is an object with metadata that may vary over time, for example, position.

Timing (timestamp) - Object audio metadata can be thought of a series of metadata events at discrete times throughout a program. The timestamp indicates when a new metadata event takes effect. Each metadata event can have, for example, updates to the position, width, or zone metadata fields.

Object Position - The position of each dynamic object is specified using three-dimensional coordinates within a normalized, rectangular room. The position is required to render an object with a high degree of spatial accuracy.

Zone / Elevation mask - The zone and elevation mask metadata fields describe which speakers, either on the listener plane or height plane of the playback environment shall be enabled or disabled during rendering for a specific object. Each speaker in the playback environment can belong to either the screen, sides, backs or ceiling zones. The mask metadata instructs the renderer to ignore speakers belonging to a given zone for rendering. For instance, to perform a front to back panning motion, it might be desirable to disable speakers on the side wall. It might also be useful to limit the spread of a wide object to the two-dimensional surround plane by disabling the elevation zone mask. Otherwise, objects are spread uniformly in three-dimensional; and will fire ceiling speakers. Finally, masking the screen would let an overhead object be rendered only by surround speakers for configurations that do not comprise ceiling channels. As such, zone mask is a form of conditional rendering metadata.

Object Width - Object width specifies the amount of spread to be applied to an object. When applied, object width increases the number of speakers used to render a particular object and creates the impression of a spatially wide source as opposed to a point source. By default, object width is isotropic and three-dimensional unless zone masking metadata is used.

Object Snap - The object *snap* field instructs the renderer to be reproduced a single loudspeaker. When object snap is used, the loudspeaker chosen to reproduce the object is typically the one closest to the original position of the object. The snap functionality is used to maximize timbral accuracy during playback. A common usecase for the snap-to-speaker parameter is to create near-screen/wide pairs of objects along the side walls by using objects which snap to the proscenium speakers (e.g. in cinema exhibition). This is particularly useful for music elements, e.g. to extend the orchestra beyond the screen. When re-rendered to sparser speaker configurations (e.g., legacy 5.1 or 7.1), these elements will be automatically snapped to left/right screen channels. Another use of the snap metadata is to create ``virtual channels", for instance to re-position the outputs of legacy multichannel reverberation plug-ins in 3D.

Object Divergence - *Divergence* is a common mixing technique used in broadcast applications. It is typically used to spread a Center channel signal (for example, primary dialogue) across the speakers in the screen plane instead of direct rendering to the center

speaker. The spread of the Center channel signal can range from all center (full convergence), through equal level in Left, Right, and Center speakers, to full divergence where all the energy is in the Left and Right speakers with none in Center speaker. Regardless of how the center signal is spread, full convergence or full divergence, the spatial image of the center signal remains consistent. The object divergence field controls the amount of direct rendering of the object compared with the rendering of two virtual sources spaced equidistantly to the left and right of the original object using identical audio. At full convergence, the object is directly rendered as it would be normally. At full divergence the object is reproduced by rendering the two virtual sources.

Number of Presentations - The *number of presentations* parameter indicates the number of presentations represented in each AC-4 frame. This parameter can be updated on a frame basis for supporting dynamic insertion/deletion of substreams during distribution and/or within consumer playback devices.

Default Presentation - The *default presentation* parameter identifies which presentation shall be utilized in consumer playback devices that do not support personalization.

Number of Substreams per Presentation - The number of substreams per presentation parameter indicates the number of substreams (carrying audio essence and metadata) associated with each *presentation* indicated in the bitstream.

Substream Content Type - The *substream content type* parameter indicates the type of essence contained in the substream. e.g. dialogue, music, FX, etc.

Substream Language - The *substream language* parameter indicates the language the substream contains or is to be associated with during rendering playback.

Gain Value - The *gain value* parameter expresses a gain to be applied during rendering for normalization and/or when mixing with associated substreams.

Additional (Non-Essential) Object Metadata

Additional metadata critical to ensure proper rendering of objects and provide sufficient artistic control include:

- Object not active
- Object priority
- Screen scaling
- Object “is mutable”
- Minimum & Maximum gain values
- Positional constraints

Object Not Active - An object may not be “on” or active throughout the lifetime of a program. The object not active field indicates that the object’s audio essence is silent or should not be rendered. The object not active field can be used by encoders, decoders, and renderers to reduce processing complexity for objects that are inactive.

Object Priority - Each object has an associated ranking of importance, object priority. The priority is not an absolute ranking of objects but rather relative a normalized value. The higher the object priority value, the more important that object’s contribution to the content is. The object priority field can be used by encoders, decoders, and renderers to reduce processing complexity, bit rate allocation, or rendering quality for objects with lower priority in constrained systems.

Screen Scaling - In cinema auditoriums, the Left and Right speakers are almost always aligned with the left and right edges of the screen. However, in consumer playback environments the position of the Left and

Right speakers can vary greatly from being adjacent to the screen to being wider, sometimes much wider, than the screen. As a result audio and visual events that are co-located on the edge of the cinema screen may not be aligned on consumer screens. To address this problem, screen scaling metadata is needed to help a renderer scale the X- and Z-positions of objects that appear onscreen. In order to work correctly, several pieces of information are required. As such screen-scaling metadata encompasses several metadata fields including:

- Master screen size ratio - The master screen size ratio represents the ratio of the screen width to the distance between the Left and Right speakers in the mastering studio.

- Use Screen Reference - The Use Screen Reference field indicates that a renderer should use the screen size as the position reference rather than the room dimension (or speaker positions) when panning the associated object. The X- and Z-positions should be scaled according to playback screen size. This will ensure that visual and audio images are aligned.

Object "is mutable" - The *object is mutable* parameter indicates whether the audio object can be completely muted during playback – under creative control.

Minimum & Maximum Gain - The *minimum and maximum gain* parameter indicates (and limits) the amount of boost/cut a consumer is able to control (in terms of level) when rendering the associated object during playback.

Positional Constraints - The *positional constraint* parameter indicates (and limits) the range of spatial re-positioning a consumer is able to control when rendering the associated object during playback.

Intelligent Loudness Metadata

The following discussion highlights the essential loudness-related metadata parameters supported the AC-4 system. Intelligent Loudness metadata provides the foundation for enabling automatic (dynamic) bypass of cascaded (real-time or file-based) loudness and dynamic range processing commonly found throughout distribution and delivery today. Intelligent Loudness metadata is supported for both channel- and object-based audio representations.

Dialogue Normalization Level – This parameter indicates how far the average dialogue level is below 0 LKFS.

Dialogue Channel – This parameter indicates whether Left, Right and/or Center channels of the substream contain dialogue.

Loudness Practice Type - This parameter indicates which recommended practice was followed when the content was authored or corrected. For example, a value of "0x1" indicates the content author (or automated normalization process) was adhering to ATSC A/85. A value of "0x2" indicates the author was adhering to EBU R 128. A special value, "0x0" signifies that the loudness recommended practice type is not indicated.

Loudness Correction Dialogue Gating Flag - This parameter indicates whether or not dialogue gating was used when the content was authored or corrected.

Dialogue Gating Practice Type - This parameter indicates what dialogue gating practice was followed when the content was authored or corrected. This parameter is typically 0x02 – "Automated Left, Center and/or Right Channel(s)". However there are values for signaling manual selection of dialogue, as well as other channel combinations.

Loudness Correction Type - This parameter indicates whether a program was corrected using a file-based correction process, or a real-time loudness processor.

Program Loudness, Relative Gated - This parameter indicates the overall program loudness as per ITU-R BS.1770-3. In ATSC geographies, this parameter would typically be -24.0 LKFS, for short-form content as per ATSC A/85. In EBU geographies, this parameter would typically indicate -23.0 LKFS (LUFS).

Program Loudness, Speech Gated - This parameter indicates the speech-gated program loudness. In ATSC geographies, this parameter would typically be -24.0 LKFS for long-form content as per ATSC A/85.

max_loudstrm3s - This parameter indicates the maximum short-term loudness of the audio program measured according to ITU-R BS.1771.

max_truepk - This parameter indicates the maximum true peak value for the audio program measured according to ITU-R BS.1770.

loro_dmx_loud_corr - This parameter is used to calibrate the downmix loudness (if applicable), as per the Lo/Ro coefficients specified in the associated metadata and/or AC-4 bitstream, to match the original (source) program loudness.

ltrt_dmx_loud_corr - This parameter is used to calibrate the downmix loudness (if applicable), as per the Lt/Rt coefficients specified in the associated metadata and/or AC-4 bitstream, to match the original (source) program loudness.

Note regarding the loudness measurement of objects: The AC-4 system [13] supports loudness estimation and correction of both channel-based and object-based (immersive) programs utilizing an extension to the ITU-R BS.1770-3 recommendation (currently under study in ITU-R Working Party 6B). Moreover, the loudness estimation element of the system dynamically applies a spatial gain for each object or object group derived from each object's or spatial object-group's positional information carried in the program's metadata. Channel-based immersive formats are also supported. Additionally, the AC-4 system supports the carriage (and control) of program loudness at the presentation level across program interchange, distribution, and consumer delivery. This ensures any *Presentation* (constructed from one or more AC-4 substreams) available to the listener will maintain a consistent loudness.

CONCLUSION

There are few areas evolving as rapidly as consumer media. Changes in consumer behavior, market dynamics, the regulatory environment, and technology are unfolding faster than ever before.

This paper outlined the latest developments in the evolution of creating and delivering more lifelike audio experiences, flexible/efficient delivery of programming that addresses the needs of the non-native speaking, hearing-impaired and visually-impaired audiences and the most essential metadata parameters necessary to enable these experiences. Thereby meeting the needs of next generation broadcast by:

Making audio personal. Viewers will be able to tailor the experience to their preference by leveraging audio Presentations and compositional control provided by the AC-4 system. This will increase audience engagement and provide new commercial opportunities for programmers and MVPDs.

Making audio more immersive. The AC-4 system has been designed from the ground up to efficiently carry both channel-based and

object-based immersive programming. Leveraging the described object-based program building blocks and system components fosters continued innovation in device playback technologies and future-proofs content against the unpredictable evolution of playback device capabilities and environments.

ACKNOWLEDGEMENTS

The authors would like to thank the following individuals for their contributions to this and the entire body of research and development in the areas of immersive and personalized sound worldwide; Scott Norcross, Tim Carroll, Sean Richardson, Elmer Musser, Steve Silva, Roger Charlesworth, Freddie Sanchez, Sripal Mehta, Prin Boon, Jonas Roden, Mike Ward, Christophe Chabanne, and Aaron Master. We appreciate your support, dedication and shared vision for next generation audio.

REFERENCES

[1] The term "Spatial Audio" refers to audio processing and reproduction techniques that enable the perception of audio essence in space and thereby in context (e.g. with accompanying picture).

[2] C. Robinson, N. Tsingos, and S. Mehta, "Scalable format and tools to extend the possibilities of cinema audio," SMPTE Motion Imaging Journal, Nov. 2012.

[3] F. Rumsey, Spatial audio. Taylor & Francis US, 2001.

[4] R. K. Furness, "Ambisonics – an overview," in AE^{S} 8th International Conference, Washington, D.C., 1990.

[5] T. Lossius, P. Baltazar, and T. de la Hogue, "DBAP–distance-based amplitude panning," in Intl. Conf. on Computer Music (ICMC), Montreal, 2009.

[6] G. Dickins, M. Flax, A. McKeag, and D. McGrath, "Optimal 3D-speaker panning," Proceedings of the AES ^{1}6th international conference, Spatial sound reproduction, Rovaniemi, Finland, pp. 421–426, April 1999.

[7] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," J. of the Audio Engineering Society, vol. 45, no. 6, pp. 456–466, June 1997.

[8] D. Kostadinov, J. D. Reiss, and V. Mladenov, "Evaluation of distance based amplitude panning for spatial audio." in Proc. of ICASSP 2010, pp. 285–288, 2010,

[9] N. Tsingos, C. Q. Robinson, D. P. Darcy, and P. A. Crum, "Evaluation of panning algorithms for theatrical applications," Proceedings of the of panning algorithms for theatrical applications," Proceedings of th^{e} 2nd Intl. Conf. on Spatial Audio (ICSA), Erlangen, Germany, February 2014.

[10] ITU-R BS.2051. Advanced sound system for programme production. http://www.itu.int/rec/R-REC-BS.2051-0-201402-I

[11] C. Robinson, N. Tsingos, Cinematic Sound Scene Description and Rendering Control, SMPTE 2014 Annual Technical Conference

[12] J. Riedmiller, S. Mehta, N. Tsingos, P. Boon, "Immersive & Personalized Audio: A Practical System for Enabling Interchange, Distribution and Delivery of Next Generation Audio Experiences", SMPTE 2014 Annual Technical Conference

[13] ETSI TS 103 190, 'Digital Audio Compression (AC-4) Standard'

[14] Dolby Atmos Specifications http://www.dolby.com/us/en/technologies/dolby-atmos/dolby-atmos-specifications.pdf

[15] Dolby Atmos
http://www.dolby.com/us/en/technologies/dolby-atmos.html