Audio Engineering Society

Convention Paper 10227

Presented at the AES 158th Convention
2025 May 22–24, Warsaw, Poland

This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Acoustic Objects: bridging immersive audio creation and distribution systems

Jean-Marc Jot[1], Thibaut Carpentier[2], and Olivier Warusfel[2]

[1]*Virtuel Works LLC, Aptos, CA, USA*
[2]*STMS Lab – CNRS, IRCAM, Sorbonne Université, Paris, France*

Correspondence should be addressed to Jean-Marc Jot (jmmjot@gmail.com)

ABSTRACT

The authoring and consumption of media experiences increasingly leverages a global set of connected frameworks and ecosystems. Once confined in separate content markets such as music, movies, video games, and virtual reality, immersive media experiences are becoming interoperable, ubiquitous, and adaptable to consumer preference, language, and listening conditions. We examine the development of object-based immersive audio technology towards unifying linear or broadcast content with nonlinear embodied experiences, while enabling flexible audio creation and playback. We analyze fragmentation obstacles hindering immersive audio technology dissemination, and advocate for the notion of *Acoustic Objects* – proposing a common spatial audio scene representation model for creating and distributing versatile, navigable sound in music, multimedia, and virtual or extended reality. An Acoustic Object is a spatial audio object for which acoustic reflection or reverberation effects can be reconstructed at the decoding/rendering end, and optionally modified by user-end customization or interaction such as remixing, karaoke, language dubbing, or 6-DoF navigation within the immersive scene.

1 The need for a common model

1.1 Types of immersive audio content

The creation and performance of immersive audio experiences is rooted in ancient theatrical and concert practices that exploit our natural hearing faculty to spatially parse multi-source soundscapes, as well as the subtle features of reflective or reverberant acoustic environments [1]. In modern times, audio technology has enabled creators to realize, record, and reproduce both realistic and imaginary artificial immersive audio experiences. These applications generally fall into two categories: linear and nonlinear.

Linear experiences involve a fixed sequence of events that can only be modified by the consumer via transport controls or predetermined versioning, including language selection. These are typical in music, movies, and podcasts. Nonlinear applications, in contrast, trigger events in response to user actions, and include video games and Extended Reality applications (VR, AR, XR) [2]. In multimodal immersive experiences, whether linear or nonlinear, creators may provide coinciding audio and visual spatial cues in order to emphasize the spectator's sense of presence [3, 4]. Modern spatial audio reprodution systems can automatically adapt to the user's playback setup (headphones or loudspeakers), as reviewed below in Section 1.1.3.

1.1.1 Aesthetics-driven linear immersive audio for music and soundtracks

Spatial audio recording and reproduction technology has evolved over several decades, from two-channel stereophonic recording to flexible immersive audio, through extended channel-based formats including surround and height channels [5]. Through this evolution, the mixing signal flow used by creators has remained consistent with the standard architecture implemented in software Digital Audio Workstations (DAWs). Its topology allows selecting among various multichannel audio panning techniques depending on the audio output format of the master bus [6, 7] (Fig. 1).

In traditional *channel-based* formats, each audio output channel is assigned to a specific loudspeaker or cluster within a predefined array layout. Recent layout-agnostic formats and low-bit-rate coding specifications, reviewed in section 1.1.3, are deployed in professional or consumer-grade media authoring software, distribution services, and playback equipment. These include *object-based* formats, illustrated in Fig. 2, and *scene-based* (a.k.a. Ambisonic) formats.

During production and mixing in a DAW, the author can adjust effect processing parameters for each source signal (Fig. 1). These parameter settings, along with their variations over time, are stored and recalled by the DAW automation system. Their combined effects in the master output can be auditioned during creation and reflected in the final export. When exporting to an object-based file format, the DAW converts these parameters into per-object metadata that encode intended spatial positions, irrespective of the playback configuration. At the consumer end, a functionally equivalent rendering engine can reproduce the immersive audio scene on a flexible loudspeaker setup or headphones (potentially with head-movement tracking).

1.1.2 Physics-driven interactive immersive audio for gaming and XR experiences

In nonlinear application scenarios like video or VR games, the end user assumes the role of a virtual listener within an acoustic environment containing one or more virtual sound sources, which may include voices of characters or avatars of remote users. Interaction may simulate six-degree-of-freedom (6-DoF) movement of the listener and/or sound sources, including translation and three-degree-of-freedom (3-DoF) rotations. The

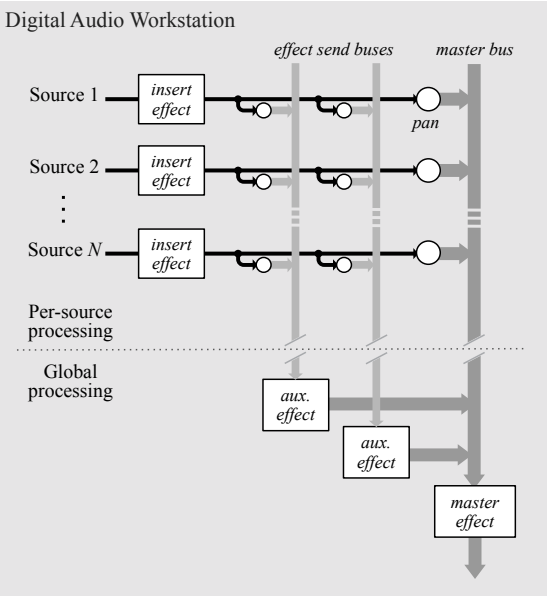


Fig. 1: Generic DAW signal flow. The topology is consistent regardless of the master output format, from two-channel stereo to modern immersive sound formats. It allows panning and mixing of multiple source tracks, with optional per source insert effects, along with global processing including shared master and auxiliary effects.

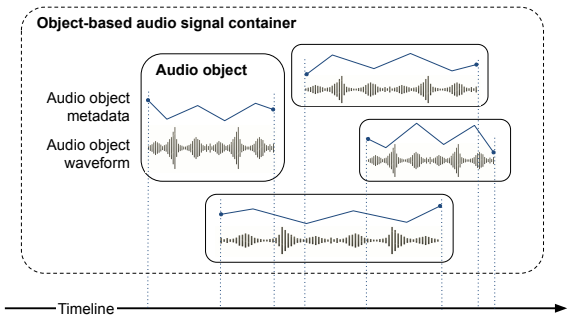


Fig. 2: Illustration of immersive audio signal format for encoding multiple virtual sound sources as audio objects. Typically, each audio object is represented by a single-channel audio waveform signal accompanied by metadata specifying its intended position and loudness.

program performs real-time spatial audio rendering, similar to a DAW mixing engine (see Fig. 1), with the output format determined at playback according to the user's headphone or loudspeaker setup.

Unlike linear playback systems, an interactive audio playback engine must obey a format-agnostic 6-DoF rendering Application Programming Interface (API) to trigger the playback of audio samples and control processing and mixing operations during runtime. In addition to per-object positional and loudness parameters, this API should specify reverberation effects, as well as sound source directivity and orientation parameters to support 6-DoF navigation of the listener and source motion. In the late 1990s, game audio APIs introduced the simulation of per-object distance, orientation, directivity, and sound muffling effects by virtual acoustic obstacles [8, 9, 10]. Later extensions allowed for simulating virtual room acoustics in more complex reflective or reverberant environments [11, 12].

In video games and XR, the audio rendering parameters are dynamically calculated by a numerical propagation model based on the virtual acoustic environment's physical and geometric description, including listener and sound source positions [13, 14, 15]. Tied to a virtual 3-D world, this model helps developers maintain audio-visual correspondence while affording 6-DoF movement for sound sources and for the listener. Propagation modeling includes reverberation evaluation and acoustic diffraction effects based on source and listener positions, as well as obstacle and room geometries. Some calculations are performed at runtime, while others may be precomputed for efficiency [16].

1.1.3 Flexible loudspeaker or binaural immersive audio reproduction

Support for flexible multichannel loudspeaker playback configurations was first introduced commercially in the early 2010s, to render interactive audio in computer sound systems. Later, automatic adaptation to the consumer's loudspeaker count and layout was introduced in home theater systems supporting modern immersive linear audio formats. Today, creators and developers of immersive audio experiences can expect automatic adaptation of the spatial presentation to the listener's loudspeaker or headphone configuration [17].

For linear immersive audio production and archiving, the Audio Definition Model (ADM) [18] offers a widely adopted open file format that can serve as precursor to low-bit-rate content encoding for distribution to consumers. Audio elements can be represented via any of the following methods:

- *Scene-based* coding, where the audio signal is encoded in a multichannel Ambisonic format [19], is implemented in MPEG-H [20] and IAMF [21].
- *Object-based* coding [22, 23] carries a collection of *spatial audio objects* (see Fig. 2), each assigned a fixed or time-varying position on a virtual surface centered on the listener or surrounding the audience. This option is implemented, for instance, in MPEG-H, Dolby Atmos [24], or DTS:X [25].
- *Channel-based* coding, equivalent to a set of audio objects assigned fixed positions matching industry-standard layouts [26], is the most common immersive audio encoding paradigm (supported in all of the distribution format examples cited above).

Immersive audio reproduction over flexible loudspeaker configurations is needed in home theaters and in automotive environments. It is commonly realized via channel-based panning techniques (see e.g. [6, 23]), or via Ambisonic rendering [19].

Many application scenarios require binaural reproduction over wearable devices such as earphones or glasses. In this case, a per-object positional rendering function can be realized by implementing a pair of dynamic digital filters realizing Head-Related Transfer Functions (HRTF) [27, 28, 29]. For a more computationally economic implementation of object-based binaural rendering, an intermediate panning and mixing format may be employed. In the case of Ambisonic mixing [30, 31], this approach is perceptually equivalent to the binaural rendering of a scene-based coded signal. Object-based binaural rendering methods advantageous in both fidelity and computational efficiency employ bilateral multichannel mixing, where the left and right ear signals are encoded separately (as illustrated in Section 2.1.2 and Fig. 4) [32, 33].

Binaural rendering methods aim to produce an immersive sound impression comparable to real-world listening or to reproduction over a multichannel loudspeaker system. In headphone or wearable audio playback, factors that impact the success of this illusion include HRTF customization, head-tracking, and listening room acoustic simulation [27, 34, 35]. An approach to restoring externalization and front/back discrimination in headphone immersive audio is introduced in [36]. Immersive binaural rendering can also be achieved via loudspeaker arrays or soundbars, in desktop or near-field personal configurations, by combining cross-talk cancellation processing, beamforming, and ear position tracking [37, 38, 39].

1.2 Limitations and fragmentation challenges

Despite remarkable advances in the ability to create and commercialize compelling immersive music recordings and cinema or game soundtracks, the democratization and success of immersive audio technology is hindered by distribution ecosystem fragmentation and content creation constraints.

1.2.1 Creation: facilitating aesthetically motivated immersive sound design

In authoring linear or nonlinear immersive audio experiences, with or without accompanying visuals, creators face limitations that impact artistic expression and workflow:

- Plausibility vs. Physical Modeling: Designing reverberation effects based on room geometry and wall materials can be artistically restrictive, even with precise acoustic simulation tools. In traditional music and soundtrack production, sound designers may wish to apply aesthetically driven effects, without a constraint of visual correspondence or authenticity of acoustic environment simulation [40].
- Democratization: DAW software was initially designed for mixing and mastering in channel-based formats. Creating object-based linear immersive audio often demands new workflows and DAW configuration to align with modern formats. Although immersive audio is in theory amenable to monitoring via headphones or low-cost loudspeaker arrays, it is often observed that the ability to critically monitor and qualify immersive audio creations involves substantial investment in studio equipment, including multichannel loudspeaker and amplification apparatus. There is a need to facilitate and democratize the creation of linear immersive audio content, towards making it as familiar and accessible as mixing and mastering in conventional stereo formats.

1.2.2 Deployment: unlocking immersive audio experience ecosystems

In gaming or interactive XR applications, as discussed in Section 1.1.2, creation and playback systems must render sound scenes where the position and orientation of the listener or of sound sources are determined during runtime rather than at the authoring stage. The predictability of the end-user quality of experience is facilitated by the fact that consumer platforms (from companies like Sony, Meta, Microsoft or HTC) are closely tied to application authoring software provided by the same vendors. In contrast, for linear applications like music or movies, consumers expect any playback device, such as a TV set, to function independently of the authoring tool employed to produce the soundtrack.

As reviewed in Section 1.1, both interactive and linear immersive audio technologies can leverage an object-based representation of the audio scene delivered to the listener. This technology advancement allows for the distribution of immersive audio content that automatically adapts to the consumer's playback setup, including their specific loudspeaker layout. However, besides this commmonality, fundamental differences and limitations have impeded the adoption and dissemination of immersive audio content:

- Interoperability. Immersive audio content markets and ecosystems are fragmented between linear and nonlinear media, and across several incompatible interactive media authoring and consumption platforms commercialized by competing companies (e.g. Apple, Microsoft, Sony or Meta).
- Personalization. In modern object-based linear immersive audio recording formats, reviewed in Section 1.1.3, end-user interaction or customization (such as remixing, karaoke, or character language replacement) is in effect restricted because, in standard practice, immersive reflections and reverberation sounds are summed up into a set of additional recorded audio waveforms (commonly referred to as the immersive recording's "bed") [22].
- Navigability. In these object-based linear formats, the audio objects composing the scene are lacking sound source directivity and orientation properties, and are assumed to be localized at a fixed distance from the listening position. This contrasts with interactive audio experiences, which allow free 6-DoF movement of the virtual sound sources and/or the virtual listener.

2 *Acoustic Objects*: bridging immersive audio creation and distribution systems

In this section, we review the proposed system illustrated in Figure 3, aiming to overcome the creation, democratization, and ecosystem fragmentation obstacles outlined in Section 1.2.

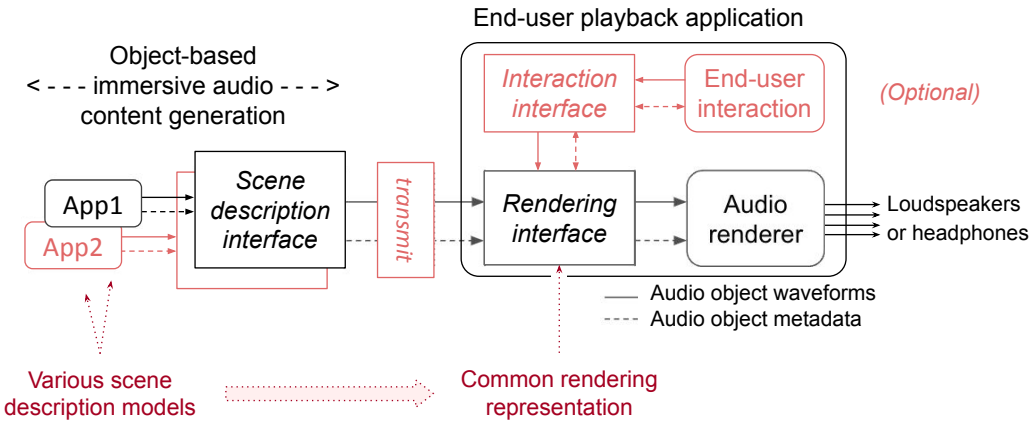


Fig. 3: Interoperable ecosystem synopsis: from application-level audio scene generation to low-level universal spatial audio rendering interface, with optional end-user interaction. Content generation paradigms may be linear or interactive, aesthetically or physically driven. For any chosen application-level content generation paradigm, a *scene description interface* translates its scene representation to an object-based rendering description compatible with the low-level *rendering interface*, thus ensuring faithful reproduction on any compatible playback device. Optionally, a bidirectional *interaction interface* allows playback-end repositioning or substitution of *Acoustic Objects*. The intermediate *transmit* function allows for the juxtaposition of several concurrent audio scenes generated by independent content generation programs.

Next, we review a general spatial audio rendering API optimized for computationally efficient physics-driven VR or XR applications, and show that it is compatible with aesthetics-driven immersive sound design. This yields a portable, application-agnostic spatial audio rendering model that is equally adept at data driven acoustic simulation and at aesthetically motivated sound design, in applications such as:

- Immersive entertainment, including games, sports, cinema, music, and virtual concerts
- Social co-presence, including virtual meetings, remote collaboration, and virtual assistants
- Virtual travel or tourism, including navigation and situational awareness
- Assisted training, education, diagnosis, surgery, design, or architecture.

Then, we introduce the notion of *Acoustic Object*, proposing to extend object-based immersive audio metadata specifications to enable application-agnostic scene representation and optional playback-end interaction in recorded or broadcast immersive audio scenes. Applications enabled by this extension include:

- In free-viewpoint VR or TV presentation of music or sport events: incorporating sound elements that coincide spatially with displayed visuals
- In cinematic movie or video playback: facilitating dialogue translation, with automatic matching of the original sound design qualities
- In music or soundtrack production: creating recordings that are readily remixable, navigable, or amenable to karaoke-like performance, to the extent authorized by content rights owners.

2.1 A universal 6-DoF spatial audio rendering interface

In [12, 41], a generic object-based rendering model is proposed for interactive 6-DoF spatial audio and multi-room reverberation, emphasizing auditory plausibility and computational efficiency. By favoring a perception-based construction of the rendered audio output, it aims for independence from application-level scene generation strategies.

2.1.1 Object-based audio rendering interface for interactive XR and gaming

The *rendering interface* encompasses the OpenAL EFX feature set [11, 12] and extends it to enable efficient per-object parametric control of the spatialization of

grouped early reflections. In order to meet the requirements of interactive gaming and XR scenarios, it allows for multiple virtual sound sources to be seamlessly instantiated by applications at rendering time, and assigned dynamic position and orientation within a navigable multi-room acoustic environment [12]. Each sound source is represented as an audio object characterized by its frequency-dependent radiation pattern (a catalog of acoustic source directivity properties can be found, for instance, in [42]).

An update in the position or orientation of the virtual listener or of a sound source triggers updates of low-level renderer parameters. The acoustic environment may be managed dynamically as a navigable set of virtual rooms, each characterized by a "reverberation preset" that encodes its response to an omnidirectional sound source located at a reference distance. This encoding incorporates the notion of *reverberation fingerprint* of a room, which affords a data-efficient characterization of the perceptually relevant attributes of acoustic reverberation that are independent of source or receiver properties and positions in the room: the reverberation fingerprint is simply specified by the volume of the room and its frequency-dependent reverberation decay time [43].

In music creation and performance scenarios, the spatial audio scene composition conceived by the author or sound designer may be most effectively represented in the perceptual domain [40]. An example of this approach is the scene description paradigm implemented in IRCAM `Spat`, described in [44]. In the following, we demonstrate the compatibility between such a perceptually based paradigm and the generic spatial audio renderer considered here.

2.1.2 Computationally efficient multi-room 6-DoF immersive audio renderer

Fig. 4 displays the architecture of a computationally efficient spatial audio rendering engine compatible with flexible loudspeaker-based or binaural playback configurations, previously described in [12, 41]. Its control interface presents the following features:

- Per-source processing: rendering of direct sound, acoustic reflections and reverberation sound field for arbitrarily complex audio scenes comprising multiple audio objects, including both point sources and spatially extended sound sources, with minimal incremental computation cost per audio object
- Effect processing: a shared set of *effect slots* that can include several parallel reverberators to simulate complex acoustic environments, and a spatialized reflections synthesis module (on the plausibility of early reflections simulation, see [28, 45, 46])
- Anisotropic reverberation: reverberator output *pan* modules enable directionally weighted or spatially diffuse rendering of the acoustic reverberation emanating from the room in which the virtual listener is located, as well as several adjacent virtual rooms simulated by additional reverberators.

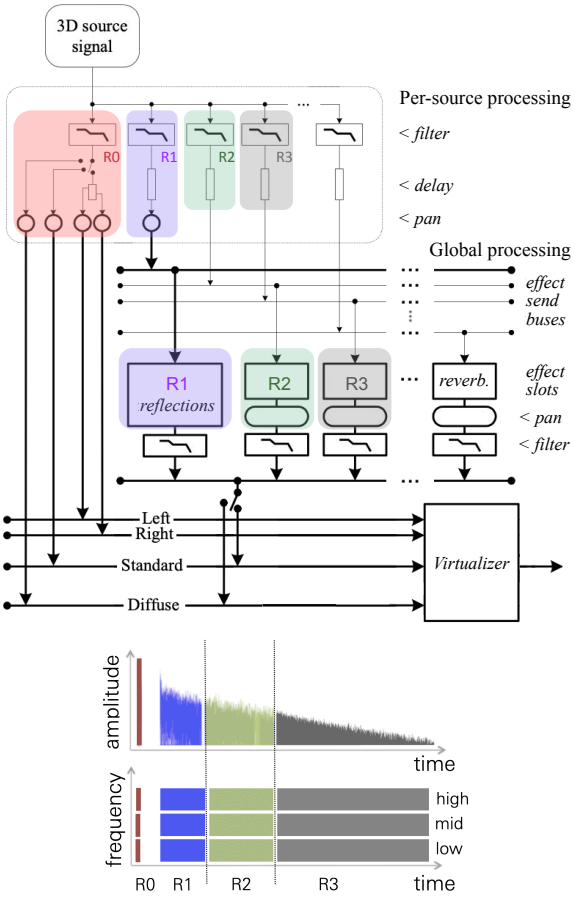


Fig. 4: Architecture of a generic immersive audio renderer for computationally efficient multi-source/multi-room object-based spatial audio over loudspeakers or headphones (adapted from [44]). Thicker lines indicate multichannel audio signal paths or processing functions. Labels (R0...R3) illustrate a configuration enabling per-source segmented reverberation rendering.

The output *Virtualizer* function can support various immersive loudspeaker or headphone playback modes, complementing the positional audio panning methods employed in the per-source direct-sound *pan* modules:

- *Standard* channel- or scene-based methods [47]
- *Left/Right* bilateral multichannel methods [32, 33]
- *Diffuse* panning dedicated to the rendering of spatially extended sounds [32, 48].

2.2 Adaptation to aesthetically-motivated immersive audio creation

2.2.1 Perceptually-based renderer configuration and control model

Fig. 4 illustrates a configuration of the generic audio renderer where reflections and reverberation processing resources are shared between audio objects assigned to the same room. The effect slots labeled R1, R2, and R3 each contain a reflections or reverberation processor configured to simulate a segment of the room impulse response. The per-source processing block adjusts onset time and spectral correction for each of the early and later reflections and reverberation segments. Details on a realization compatible with IRCAM `Spat` are described in [44].

In order to match auditory perspective, the low-level rendering interface must be listener-centric: object position and orientation coordinates are defined relative to the listener's coordinates and head orientation. Although the notion of "room" may be exposed at the application level, the renderer sees its task merely as a superposition of sound components (direct sound, reflections, reverberation), each specified via a minimal set of rendering properties (see Table 1). Similarly, the source's orientation and directivity parameters are not required in the low-level rendering interface.

2.2.2 *Spatial Audio Object Workstation*: facilitating linear immersive audio creation

We define a *Spatial Audio Object Workstation* (SAOW) as a linear immersive audio authoring tool that exposes the control parameters of the generic renderer [44]. As shown in Fig. 5, the renderer's signal flow can be mapped to the familiar topology of a standard DAW (Fig. 1). The *effect send bus* is configured automatically according to the configuration of the *effect slots*, thus abstracting audio signal routing configuration to render object-based reflections and reverberation.

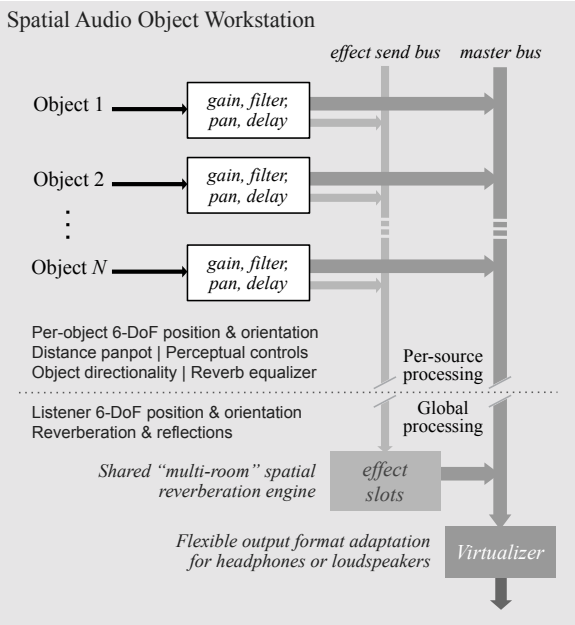


Fig. 5: Realizing a *Spatial Audio Object Workstation* by mapping the generic renderer of Fig. 4 to the standard DAW signal flow of Fig. 1. The per-object processing functions and the signal formats of the internal multi-channel summing buses (the *effect send bus* and the *master bus*) are apparent in Fig. 4 (adapted from [44]).

For the mixing engineer, operating a SAOW is consistent with the familiar practices and workflows employed for producing stereo or immersive audio recordings in channel-based or scene-based audio formats (defined in Section 1.1.3). Additionally, a SAOW exposes sound design functions surpassing traditional audio production tools:

- A per-source distance/depth control may be realized by a "distance panpot" which has the effect of simultaneously adjusting the levels and/or onset times of R1, R2 and R3 for each audio object [44, 49]
- Per-source on-axis spectral correction and frequency-dependent directivity pattern, which affect the direct-sound frequency response according to the source's orientation relative to the position of the listener
- Per-source diffuse-field spectral correction, which enables correcting the spectral contour of the reverberation and reflections individually for each sound source (simulating the diffuse-field transfer function of this source, as defined in [50]).

2.3 *Acoustic Objects*: unifying the immersive audio ecosystem

In this section, we introduce the notion of *Acoustic Object*, defined as a spatial audio object whose metadata describe its reverberation and reflection properties, thereby enabling rich capabilities of interaction and modification of the virtual acoustic scene by the consumer. We begin by reviewing prior research and approaches to the object-based description of environmental acoustic scenes.

2.3.1 Audio scene metadata interoperability requirements

Within the ecosystem represented in Fig. 3, it is advantageous to adopt a vendor-neutral *rendering interface* facilitating the interoperability of devices and applications, and to define a media encoding format so that a scene can be displayed by a playback application agnostic to the scene generation program (which may be, for instance, a DAW or a video game application). A high-level *scene description interface* converts an application-specific scene representation to a vendor-neutral specification compatible with the common low-level *rendering interface*: a set of audio object waveform, stream, or synthesizer references, along with metadata that collectively drive the spatial audio processing operations performed by the audio renderer.

For each spatial audio object, its *essential metadata* must determine the parameters required for direct-sound rendering in free field: source directivity pattern (frequency-dependent), size or shape, position and orientation coordinates. In interactive use cases, such as gaming or augmented reality, some of this information may be unknown until playback time, and rendering algorithm parameters must therefore be updated dynamically.

As for the additional metadata that drive the rendering of reverberation and reflections, two approaches have been proposed:

- *Object-based:* incorporating reverberation processing parameters in an object's metadata, as in:
  - *MPEG-4 Advanced AudioBIFS (Perceptual Approach)*, with a parametrization based on IRCAM `Spat` [51]
  - *RSAO (Reverberant Spatial Audio Object)*, which provides a set of discrete reflections combined with a parametric reverberation decay tail [52].
- *Environment-based:* a separate metadata interface describes the acoustic environment, as in:
  - *MPEG-4 Advanced AudioBIFS (Physical Approach)*, from the DIVA research project [53, 51]
  - *OpenAL Effects Extension*, an audio programming interface providing immersive reverberation as a shared parametric effect processing option [11]
  - *MPEG-I*, which includes the constructs *Acoustic Environment* and *Geometry*, separate from *Sources* [54].

In the latter approach, the environment may be described in architectural terms, including boundary and obstacle geometry, openings and materials. From such geometrical and physical information, combined with each object's essential metadata, an acoustic propagation computing engine can derive the values of the rendering parameters (see e.g. [53, 55, 16, 15]). A common method consists of partitioning the acoustic environment into a set of rooms, so that a playback-time update in the localization coordinates of an object will inform its assignment to one of these rooms and determine its contribution to effect send buses [11, 12].

2.3.2 End-user interaction with Acoustic Objects

Optionally, the playback application of Fig. 3 may enable end-user interaction with the presented immersive audio scene via the *interaction interface*, which can support bidirectional functionality:

- Audio object modification or generation (analogous to the *scene description interface* function)
- Retrieving the values of the metadata associated with an incoming audio object to display or represent its properties in the end-user interaction interface (shown at the right-hand side of Fig. 3).

In defining a universal Acoustic Object rendering interface, desirable objectives include:

- Affording maximum flexibility in per-object sound design intent by the content creator
- Ensuring faithful reproduction of the intended spatial audio scenes at the playback end
- Letting the end user take advantage of the interactivity and personalization benefits enabled by the Acoustic Object representation, within limits authorized by the content creator
- Enabling content creators and playback application developers to select among several possible levels of end-user interaction and implementation complexity.

In the rendering interface, the metadata set associated with an Acoustic Object includes acoustic reverberation properties in addition to the per-object loudness and angular position properties already supplied in existing object-based immersive audio formats (see Section 1.1.3). In Table 1, we categorize the reverberation properties of an Acoustic Object into a hierarchy of *functionality profiles*, exposing increasing levels of end-user interaction.

2.3.3 Dynamic Acoustic Object renderer and synchronized metadata

Upon reception (or termination) of an incoming audio object, the *rendering interface* function of Fig. 3 can instruct the Audio Renderer to automatically configure its audio processing signal flow (see Fig. 4), based on interpreting per-object metadata:

- Instantiate (or deallocate) the *per-source processing* module associated with this audio object;
- If this audio object is an Acoustic Object:
  - Load (or free up) *effect slot(s)* necessary to render the reflections and/or reverberation called for according to the parameters received via the incoming per-object metadata (see Table 1);
  - Load (or free up) the corresponding *effect send buses*, along with their connections with the *per-source processing* module associated with this audio object.

The end-user playback application of Fig. 3 is compatible with the playback of legacy stereo, as well as channel-based or object-based immersive audio formats (see Section 1.1.3). The transmission of one or more Acoustic Objects is optional, and involves extending the description of each of these audio objects with a synchronized metadata stream that specifies its reflections and reverberation properties. In practice, this solution can be implemented, for instance, as an extension of the ADM format [18].

3 Conclusion: unifying immersive audio experiences

In the modern connected media ecosystem, it is essential to adopt interoperable transmission standards such that independent content generation applications can seamlessly deliver sounds presented to end users employing generic spatial audio playback devices.

Basic	Minimum Acoustic Object metadata set to ensure faithful rendering of this object (also allows gain scaling, removal, substitution, or 3-DoF repositioning of this object):
	Gain factor to be applied to this object.
	Position and *Extent* of the sound source may be jointly represented by a normalized *divergence panning vector* (p0) [41, 44, 56]
	Reverberation response: for an Acoustic Object whose reverberation is decomposed into N response segments (denoted R1...RN, see Fig. 4), their energies and the late reverberation decay time (frequency dependent), their onset times ($t1...tN$), and their divergence panning vectors ($p1...pN$).
Remix	Additional metadata to enable and control spectral equalization for this object (useful e.g. in playback-end remixing, mashup or karaoke application scenarios):
	Spectral corrections: *on-axis* (affecting the object's direct sound) and *diffuse-field* (affecting its reverberation and reflections) [44].
Navig.	Additional metadata to enable and control playback-end 6-DoF navigation and repositioning of this object [11, 12]:
	Orientation coordinates and frequency-dependent off-axis directivity;
	Distance in "world units," and optional parameters for controlling *distance-based attenuation factors*;
	Optional parameters for controlling features such as *air absorption*, *Doppler effect* or *propagation delay*.

Table 1: Hierarchy of metadata functionality profiles in Acoustic Object encoding specification (adapted from [44]), with a reverberation response decomposed into N segments (N is typically 2 or 3, and may be freely set for each Acoustic Object). The *Basic* profile includes the minimum set of properties necessary for faithful immersive audio reproduction. The *Remix* profile enables object substitution such as language replacement, where both an audio object's waveform and its source radiation properties can be swapped, with sound design qualities preserved. The *Navig.* profile supplies additional per-object properties enabling 6-DoF repositioning of an Acoustic Object and/or 6-DoF navigation of the listener.

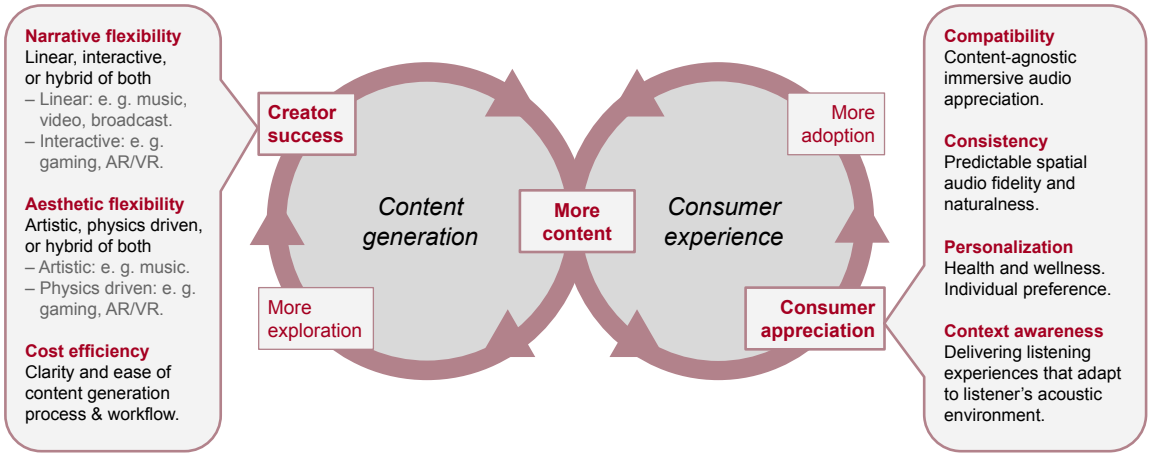


Fig. 6: Representation of the immersive audio ecosystem in the form of two interconnected "value flywheels": technology adoption is accelerated by decoupling and democratizing content generation and augmenting consumer experience, and by allowing the two industries to develop independently while supporting each other, through leveraging an interoperable content format.

By enabling the transmission and rendering of sounds in the form of *Acoustic Objects*, the end-to-end solution shown in Fig. 3 aims to mitigate the market fragmentation challenges reviewed in Section 1.2.

As illustrated in Fig. 6, the unification of linear and nonlinear platforms, as well as the integration of artistic and physics-driven sound design, facilitate the dissemination of immersive audio and music experiences and the widespread adoption of the consumer devices employed to enjoy them:

- consumer adoption is promoted by the consistently met expectation of a compelling and natural-sounding hearing experience, and by a growing content and experience catalog
- content production and experimentation are encouraged by increasing consumer demand and confidence, along with the deployment of efficient creation workflows (see Section 2.2.2).

For linear immersive audio playback, rendering Acoustic Objects requires moderate extra computation for implementing the *effect slots* and the *effect send buses* (defined in Section 2.1 and shown in Fig. 4).

3.1 A common immersive audio content model

At the content generation end, the encoding model introduced in Section 2.3 provides the option to transmit as Acoustic Objects one or more of the spatial audio objects composing an immersive scene, extending the capabilities of the existing immersive audio coding formats reviewed in Section 1.1.3. An Acoustic Object is a spatial audio object for which acoustic reflection or reverberation effects can be reconstructed at the reception end, and optionally modified by user-end customization or interaction features such as remixing, karaoke, language dubbing, or 6-DoF navigation within the immersive audio scene (see Section 2.3.2).

At the receiving end, the incoming immersive audio content is handled identically whether the narrative intent of the author is linear (e.g. a cinematic soundtrack) or interactive (e.g. a VR game soundtrack). The choice of linear or nonlinear narration is determined solely at the content origination end, as is the choice of aesthetic design paradigm (e.g. physics-driven or psychoacoustics-driven). The *Scene description interface* function associated with a content generation application (left-hand side of Fig. 3) ensures that the application-specific audio scene representation will be automatically translated to standard renderer-compatible Acoustic Object metadata.

In linear content creation, such as immersive music production, the *Spatial Audio Object Workstation* (SAOW), introduced in Section 2.2.2, alleviates the "immersive audio barrier to entry" evoked in Section 1.2.1 by facilitating object-based audio production through an intuitive extension of familiar DAW signal flows.

3.2 Unlocking the immersive audio experience ecosystem

In immersive audio playback systems – including home, automotive, mobile or wearable consumer devices – the capability to decode and render Acoustic Objects enables device optimizations and user experience enhancements:

- Interoperability: The playback function is agnostic to the type of incoming immersive audio content. It can be optimized to run permanently using dedicated computation resources, and render audio objects generated by several concurrent programs.
- Consistency: As a result, the timbral and spatial audio quality delivered to the consumer can be consistently personalized and predictable, irrespective of the type of content or experience (e.g. a motion picture soundtrack, an immersive music recording, or a VR meeting or concert).
- Personalization: For any sound encoded as an Acoustic Object, a rich set of customization or interaction features may be authorized by the content creator, ranging from hearing enhancements to musical remix, language substitution or 6-DoF navigation (as outlined in Section 2.3.2 and Table 1).
- Context awareness: In binaural reproduction scenarios, using for instance earphones or wearables, one important type of consumer-end personalization is the dynamic simulation of the user's local environment *reverberation fingerprint*, enabling natural-sounding augmented reality audio experiences (see section 2.1.1). In loudspeaker reproduction, on the other hand, acoustic objects can be reproduced faithfully by accounting and compensating for the user's local environment reverberation (see [44, 57]).

The development of multichannel audio formats amenable to flexible headphone or loudspeaker configuration has played a critical role to enable the commercial dissemination of linear immersive audio content, including music and cinematic soundtracks. By further extending object-based immersive audio formats to support the optional transmission of *Acoustic Objects*, the audio industry can overcome a persistent fragmentation obstacle: ensure that any immersive audio experience, interactive or linear (including broadcast or streaming), aesthetics-driven or physics-driven, can be reproduced seamlessly on any consumer device – while allowing end-user personalization and interaction options including hearing enhancement, automatic language dubbing, 6-DoF navigation, or creative interaction.

Acknowledgments

The authors would like to thank Agnieszka Roginska, Alan Silva, David Pearl, Earl Vickers, Luiz Zanardo, Paul Geluso, Peter Otto and the anonymous reviewers for thoughtful comments and feedback on aspects covered in this paper.

References

[1] Boren, B., "History of 3D Sound," in [5], chapter 2, 2018.

[2] Llewellyn, G. and Paterson, J., "Towards 6DOF: 3D audio for virtual, augmented, and mixed realities," in J. Paterson and H. Lee, editors, *3D Audio*, Routledge, New York, 2021.

[3] Schubert, T., Friedmann, F., and Regenbrecht, H., "The Experience of Presence: Factor Analytic Insights," *Presence: Teleoper. Virtual Environ.*, 10(3), 2001.

[4] André, C. R., Corteel, É., Embrechts, J.-J., Verly, J. G., and Katz, B. F., "Subjective evaluation of the audiovisual spatial congruence in the case of stereoscopic-3D video and Wave Field Synthesis," *Int. J. Human-Computer Studies*, 72(1), 2014.

[5] Roginska, A. and Geluso, P., editors, *Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio*, Routledge, 2018.

[6] Pulkki, V., "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, 45(6), 1997.

[7] Zhang, W., Samarasinghe, P. N., Chen, H., and Abhayapala, T. D., "Surround by Sound: A Review of Spatial Audio Recording and Reproduction," *Applied Sciences*, 7(532), 2017.

[8] IA-SIG, *3D Audio Rendering and Evaluation Guidelines*, MIDI Manufacturers Assoc., 1998.

[9] IA-SIG, *3D Audio Rendering Guidelines - Level 2.0*, MIDI Manufacturers Assoc., 1999.

[10] Hiebert, G., *OpenAL 1.1 Specification and Reference*, Creative Labs, 2005.

[11] Peacock, D., Harrison, P., D'Orta, A., Carpentier, V., and Cooper, E., *OpenAL Effects Extension Guide*, Creative Labs, 2006.

[12] Jot, J.-M., Audfray, R., Hertensteiner, M., and Schmidt, B., "Rendering Spatial Sound for Interoperable Experiences in the Audio Metaverse," in *Proc. Int. Conf. Immersive and 3D Audio (I3DA)*, 2021.

[13] Vorländer, M., *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, Springer, 2008.

[14] Savioja, L. and Svensson, U. P., "Overview of geometrical room acoustic modeling techniques," *J. Acoust. Soc. Am.*, 138(2), 2015.

[15] Pind, F., Einarsson, J., Guojonsson, S., Cosnefroy, M., Pedersen, J., Stefansson, J., and Milo, A., "A Novel Wave-Based Virtual Acoustics and Spatial Audio Framework," in *Proc. AES AVAR Conf.*, 2022.

[16] Liu, S. and Manocha, D., *Sound Synthesis, Propagation, and Rendering*, Springer, 2022.

[17] Rumsey, F., "Immersive Audio: Objects, Mixing, and Rendering," *J. Audio Eng. Soc*, 64(7/8), 2016.

[18] Füg, S., Marston, D., and Norcross, S., "The Audio Definition Model – A Flexible Standardised Representation for Next Generation Audio Content in Broadcasting and Beyond," in *Proc. 141st AES Convention*, 2016.

[19] Zotter, F. and Frank, M., *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, Springer, 2019.

[20] Herre, J., Hilpert, J., Kuntz, A., and Plogsties, J., "MPEG-H Audio—The New Standard for Universal Spatial/3D Audio Coding," *J. Audio Eng. Soc*, 62(12), 2015.

[21] Hwang, S. and Lim, F., "Immersive Audio Model and Formats, v1.1.0," 2024.

[22] Riedmiller, J., Mehta, S., Tsingos, N., and Boon, P., "Immersive and Personalized Audio: A Practical System for Enabling Interchange, Distribution, and Delivery of Next-Generation Audio Experiences," *SMPTE Motion Imaging J.*, 124(5), 2015.

[23] Tsingos, N., "Object-Based Audio," in [5], chapter 8, 2018.

[24] Robinson, C. Q., Mehta, S., and Tsingos, N., "Scalable Format and Tools to Extend the Possibilities of Cinema Audio," *SMPTE Motion Imaging J.*, 121(8), 2012.

[25] Jot, J.-M. and Fejzo, Z., "Beyond Surround Sound - Creation, Coding and Reproduction of 3-D Audio Soundtracks," in *Proc. 131st AES Convention*, 2011.

[26] ITU, "Recommendation BS.2051-3 – Advanced sound system for programme production," 2022.

[27] Roginska, A., "Binaural Audio Through Headphones," in [5], chapter 4, 2018.

[28] Jot, J.-M., Larcher, V., and Warusfel, O., "Digital Signal Processing Issues in the Context of Binaural and Transaural Stereophony," in *Proc. 98th AES Convention*, 1995.

[29] Zotkin, D., Duraiswami, R., and Davis, L., "Rendering localized spatial audio in a virtual auditory space," *IEEE Trans. Multimedia*, 6(4), 2004.

[30] Noisternig, M., Sontacchi, A., Musil, T., and Holdrich, R., "A 3D Ambisonic Based Binaural Sound Reproduction System," in *Proc. AES 24th Int. Conf. Multichannel Audio*, 2003.

[31] Rafaely, B., Tourbabin, V., Habets, E., Ben-Hur, Z., Lee, H., Gamper, H., Arbel, L., Birnie, L., Abhayapala, T., and Samarasinghe, P., "Spatial audio signal processing for binaural reproduction of recorded acoustic scenes - review and challenges," *Acta Acustica*, 6(47), 2022.

[32] Jot, J.-M., Philp, A., and Walsh, M., "Binaural Simulation of Complex Acoustic Scenes for Interactive Audio," in *Proc. 121st AES Convention*, 2006.

[33] Ben-Hur, Z., Alon, D. L., Mehra, R., and Rafaely, B., "Binaural Reproduction Based on Bilateral Ambisonics and Ear-Aligned HRTFs," *IEEE Trans. Audio, Speech, and Language Processing*, 29, 2021.

[34] Best, V., Baumgartner, R., Lavandier, M., Majdak, P., and Kopčo, N., "Sound Externalization: A Review of Recent Research," *Trends in Hearing*, 24, 2020.

[35] Frankenbach, C., Herzog, C., Seitz, J., and Fels, J., "Evaluating Localization Accuracy of Dynamic Binaural Synthesis: Differences Between Individual, Estimated, and Generic HRTFs," *J. Audio Eng. Soc.*, 73, 2025.

[36] Jot, J.-M., Lukin, A., and Landschoot, C., "Binaural Externalization Processing – from Stereo to Object-Based Audio," in *Proc. 153rd AES Convention*, 2022.

[37] Choueiri, E., "Binaural Audio Through Loudspeakers," in [5], chapter 5, 2018.

[38] Gálvez, M. F. S., Menzies, D., and Fazi, F. M., "Dynamic Audio Reproduction with Linear Loudspeaker Arrays," *J. Audio Eng. Soc*, 67(4), 2019.

[39] Gardner, W., *3-D Audio Using Loudspeakers*, Ph.D. thesis, Massachusetts Institute of Technology, 1997.

[40] Filimowicz, M., editor, *Foundations in Sound Design for Linear Media – A Multidisciplinary Approach*, Routledge, 2019.

[41] Jot, J.-M., "Interactive 3D audio rendering in flexible playback configurations," in *Proc. Asia Pacific Signal and Information Processing Assoc. Annual Summit*, 2012.

[42] Ackermann, D., Brinkmann, F., and Weinzierl, S., "A Database with Directivities of Musical Instruments," *J. Audio Eng. Soc*, 72(3), 2024.

[43] Jot, J.-M. and Lee, K.-S., "Augmented Reality Headphone Environment Rendering," in *Proc. AES Conf. Audio for Virtual and Augmented Reality*, 2016.

[44] Jot, J.-M., Carpentier, T., and Warusfel, O., "Perceptually Motivated Spatial Audio Scene Description and Rendering for 6-DoF Immersive Music Experiences," in *Proc. 2nd International Conference on Immersive and 3D Audio (I3DA)*, 2023.

[45] Gospodarek, M., Warusfel, O., Ripollés, P., and Roginska, A., "Methodology for perceptual evaluation of plausibility with self-translation of the listener," in *Proc. AES International Conference on Audio for Virtual and Augmented Reality*, 2022.

[46] Meyer-Kahlen, N., Schlecht, S., Amengual Garí, S., and Lokki, T., "Testing Auditory Illusions in Augmented Reality: Plausibility, Transfer-Plausibility, and Authenticity," *J. Audio Eng. Soc.*, 72, 2024.

[47] Spors, S., Wierstorf, H., Raake, A., Melchior, F., Frank, M., and Zotter, F., "Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State," *Proc. IEEE*, 101(9), 2013.

[48] Potard, G. and Burnett, I., "Decorrelation Techniques for the Rendering of Apparent Sound Source Width in 3D Audio Displays," in *Proc. 7th Int. DAFx Conference*, 2004.

[49] Gerzon, M. A., "The Design of Distance Panpots," in *Proc. 92nd AES Convention*, 1992.

[50] Jot, J.-M., Cerveau, L., and Warusfel, O., "Analysis and Synthesis of Room Reverberation Based on a Statistical Time-Frequency Model," in *Proc. 103rd AES Convention*, 1997.

[51] Väänänen, R. and Huopaniemi, J., "Advanced AudioBIFS: virtual acoustics modeling in MPEG-4 scene description," *IEEE Trans. Multimedia*, 6(5), 2004.

[52] Coleman, P., Franck, A., Jackson, P. J., Hudges, R. J., Remaggi, L., and Melchior, F., "Object-Based Reverberation for Spatial Audio," *J. Audio Eng. Soc.*, 65(1/2), 2017.

[53] Savioja, L., Huopaniemi, J., Lokki, T., and Väänänen, R., "Creating Interactive Virtual Acoustic Environments," *J. Audio Eng. Soc*, 47(9), 1999.

[54] Herre, J. and Disch, S., "MPEG-I Immersive Audio – Reference Model For The Virtual/Augmented Reality Audio Standard," *J. Audio Eng. Soc.*, 71(5), 2023.

[55] Raghuvanshi, N. and Snyder, J., "Parametric Directional Coding for Precomputed Sound Propagation," *ACM Trans. Graphics*, 37(4), 2018.

[56] Gerzon, M. A., "General Metatheory of Auditory Localisation," in *Audio Eng. Soc. 92nd Convention*, 1992.

[57] Menzies, D. and Fazi, F. M., "A Perceptual Approach to Object-Based Room Correction," in *Proc. 141st AES Convention*, 2016.