Perceptually Motivated Spatial Audio Scene Description and Rendering for 6-DoF Immersive Music Experiences

Jean-Marc Jot, Thibaut Carpentier, Olivier Warusfel

Perceptually Motivated Spatial Audio Scene Description and Rendering for 6-DoF Immersive Music Experiences

Jean-Marc Jot
Virtuel Works LLC
Aptos, CA USA
jmmjot@gmail.com

Thibaut Carpentier
STMS Lab, CNRS,
IRCAM, Sorbonne Université
Paris, France
thibaut.carpentier@ircam.fr

Olivier Warusfel
STMS Lab, IRCAM,
CNRS, Sorbonne Université
Paris, France
olivier.warusfel@ircam.fr

***Abstract*—In recorded, live, virtual- or augmented-reality immersive musical experiences, the audio presentation may not be prescribed by the geometrical and physical model of a room or acoustic enclosure (as may often be expected, for instance, in architectural acoustic design and in video games or virtual reality). In previous work, a generic spatial audio rendering engine and interface were proposed, compatible with both physically- and perceptually-based representations of interactive audio scenes navigable at playback time. This parametric model builds and extends upon concepts proposed previously in spatial audio programming or media standard specifications, including OpenAL and MPEG-4. It enables prioritizing auditory plausibility over simulation exactness, facilitating low-complexity implementation even for large multi-object audio scenes, by sharing a "multi-room" acoustic reverberation and reflections processor. In this paper, we review seminal research on subjective room acoustics and the design of the perceptually based spatialization interface realized in IRCAM *Spat*. The resulting *"Spatial Audio Object Workstation"* is consistent with familiar content creation tools and workflows, and enhances their functionality with fine control of the spatial motion (including distance and orientation), directionality, presence, and reverberance of each sound source. Finally, we propose a generic immersive audio coding format as a container of *"6-DoF spatial audio objects"* to advance a future interoperable audio content and experience ecosystem.**

***Index Terms*—spatial audio, virtual acoustics, room acoustics, reverberation, object-based audio, immersive music, virtual reality, augmented reality, networked music performance.**

I. Introduction

A. Approaches to audio scene description

In interactive media such as video games or virtual reality experiences, the sonic content consists of a collection of *audio objects*, each characterized by a source audio waveform accompanied by object metadata. Combined with virtual acoustic environment parameters, the object metadata serve to guide signal processing and mixing computations executed at playback time. By tying these control data to a geometrical and physical description of a virtual 3-D world, the application developer can manage and maintain the correspondence between the audio and visual presentations, while affording six degrees of positional freedom to each of the sound sources and to the listener (or camera), as described e.g. in [1].

Fig. 1. A live immersive music experience taking place in *Espace de Projection*, IRCAM's variable acoustics concert hall (credit: Quentin Chevrier).

The sound field and propagation models adopted in physics-based audio creation tools and rendering engines can draw from abundant prior research in the fields of architectural acoustics and auralization (e.g. [2]–[6]). In order to facilitate game or virtual reality audio creation, these computational models can be complemented by methods that enable a sound designer to specify and control aesthetically motivated adjustments to the final presented mix [7] (for instance, a correction in reverberation loudness for some or all digital audio objects). Such aesthetically motivated corrections often cannot be encoded exclusively or practically in terms of physical or architectural properties of the virtual environment.

In the production of linear media such as music, movies or podcasts, traditional audio creation workflows leverage time-based multitrack editing and digital audio workstation (DAW) frameworks. Each track or "stem" can be manipulated as an individual audio object until the mix is completed, mastered and exported for distribution [8], [9]. Such tools allow creators to perform aesthetically motivated spatial sound manipulations, free from constraints of exact visual correspondence or rendition of a particular reverberant space.

B. Immersive music experiences with spatial audio objects

Recently deployed object-based multi-channel digital audio standards empower media creators to produce format-agnostic immersive audio content compatible with playback over headphones or flexible loudspeaker configurations [10]–[12]. Commercialized solutions assume a static listener (allowed free head rotation, i.e. three degrees of freedom [8]): each audio object is equivalent to a virtual loudspeaker pointing towards the listener and positioned at a fixed distance. Object metadata enable precise dynamic positioning of reverberation-free audio objects along both azimuth and elevation. Reverberation and reflections are collected and encoded into a separate set of static audio object channels, often referred to as the "bed".

In recent or previous proposals and standards [12]–[19], audio object metadata include additional descriptors for controlling the dynamic rendering of reverberation and reflections, thereby supporting the encoding of *"6-DoF spatial audio objects"* amenable to free spatial motion, suppression or substitution at playback time. This paradigm allows the creation and transmission of navigable and remixable audio scenes for gaming, virtual reality or musical applications. Virtual concert experiences may leverage spatial audio object transmission so that the audio presentation will coincide spatially with a visual presentation displayed to the freely moving spectator [20]–[22]. This opportunity also applies to augmented reality and networked music performance (where the listener is present with some of the performers or is an active participant) connecting distributed venues having matched, mismatched or controllable room acoustics (Fig. 1) [23]–[27].

C. From scene description to audio rendering

Fig. 2 illustrates a multi-application immersive audio ecosystem, differentiating application-level *scene description* vs. low-level spatial audio *rendering interface* [1]. Adopting a perceptually grounded object-based rendering interface facilitates device implementation efficiency and interoperability with applications developed by independent parties. It also brings additional developer and end-user advantages:

- Experiences that seamlessly blend physically and perceptually represented audio objects or environment properties (e.g. non-diegetic sounds in VR, as discussed in [28])
- Perceptually-based modifications of audio objects or environment properties, specified at creation time or determined at playback time, that do not lend themselves to practical control via physically represented scene modifications (e.g. a sound warping effect meant to evoke emotional state)
- Compensation of listening room reverberation in loudspeaker-based playback (addressed in section IV-A).

In section II, we briefly review a generic 6-DoF object-based renderer model previously described in [1]. In sections III and IV, we revisit the design principles and scene description concepts in IRCAM Spat, a perceptually-based immersive music creation and performance tool [29]. In section V, we introduce the *"Spatial Audio Object Workstation"* and propose a metadata specification for the encoding of immersive audio scenes as collections of 6-DoF spatial audio objects.

II. A Generic 6-DoF audio scene rendering model

In [1], 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 favouring a perception-based construction of the generated audio output, it aims for independence from application-level scene description strategies, along the principles proposed previously in [16], [30].

A. 6-DoF object-based audio rendering interface

The proposed *rendering interface* (see Fig. 2) encompasses the OpenAL EFX feature set [31] and extends it to enable efficient per-object parametric control of the spatialization of grouped early reflections (as proposed previously in [17], [32]). In order to meet the requirements of interactive gaming and virtual or augmented reality 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. Each sound source is represented as an audio object characterized by its frequency-dependent directivity pattern.

An update in the position or orientation of the virtual listener or of a sound source triggers one or more updates of low-level renderer parameters. In the case of an application-level geometrical/physical scene representation, the acoustic environment may be managed dynamically as a navigable set of virtual rooms. As proposed in [1], each one of these virtual rooms may be 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 its acoustic reverberation that are independent of source or receiver properties and positions in this room [33].

In the present paper, we focus particularly on music creation and performance scenarios. In this case, as discussed in section I-A, the spatial audio scene composition conceived by the author or sound designer may be most effectively represented in the perceptual domain (optionally but not necessarily complemented by a virtual architectural/physical description). A representative example of this approach is the perceptually based scene description paradigm implemented in IRCAM Spat, examined in section IV. In section V, we discuss the compatibility between this perceptually-based paradigm and the generic rendering interface proposed in [1], together with a corresponding audio processing architecture summarized in the next section and represented in Fig. 3.

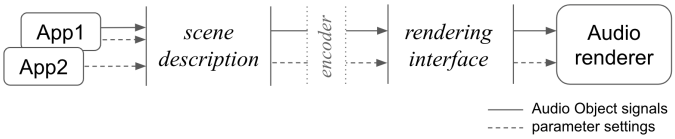

Fig. 2. Interoperable ecosystem synopsis [1]: from application-level audio *scene description* to low-level spatial audio object *rendering interface*. (The optional *encoder* function is discussed in Section V-C.)

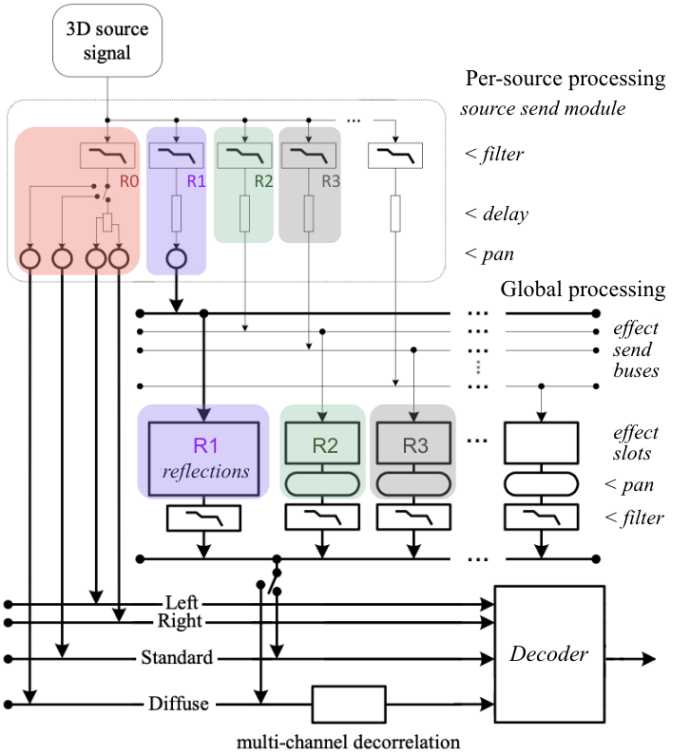

Fig. 3. Architecture of a generic 6-DoF renderer for computationally efficient multi-source/multi-room object-based spatial audio over loudspeakers or headphones. Thicker lines indicate multi-channel audio signal paths or processing functions. Color highlights illustrate the realization, described in section V-A, of a single-room renderer compatible with the Spat_OPer perceptually-based sound spatialization control interface, and equivalent to the Spat renderer algorithm described in section IV-B (Figs. 7 and 9).

B. An object-based multi-room spatial audio renderer

Fig. 3 displays the architecture of a computationally efficient spatial audio rendering engine implementing the proposed generic rendering interface model. It presents, in particular, the following advantageous features [1], [17], [32]:

- Rendering of direct sound, acoustic reflections and reverberation sound field for arbitrarily complex audio scenes comprising hundreds of audio objects, including both point sources and spatially extended sound sources
- Minimal incremental computation cost per audio object, by incorporating a shared reverberation processing engine including a grouped spatialized reflections synthesis module
- Directionally weighted or spatially diffuse (isotropic) rendering of the acoustic reverberation emanating from the room in which the virtual listener is located and from multiple adjacent virtual rooms.

The multichannel audio *Decoder* function can support various immersive loudspeaker or headphone playback modes, complementing the positional audio methods employed in the direct-sound *pan* modules: *Left/Right* bilateral multichannel methods [32], [34]; *Standard* channel-based or scene-based (Ambisonic) methods [8], [35]; *Diffuse* bus dedicated to the rendering of spatially extended sound components [17], [32].

III. Perceptually Based Scene Representation

In this section, we summarize the background, approach and results of a psycho-experimental study that informed the development of a spatial reverberation processor controlled by a perceptually based interface. The design of this tool and of the broader Spat library will be reviewed in Section IV.

A. Research on the perception of auditorium reverberation

The interest in performance venues for the enjoyment of live orchestral music, opera or theater has prompted extensive scientific research on the architectural design of auditoria and on the objective characterization and perceptual evaluation of their room acoustical quality (e.g. [36]–[39]). Following the distinction proposed in [36], one can differentiate several domains for describing the effect of room reverberation in the context of music listening, recording or production (Fig. 4):

- *Architectural*: referring to geometrical and physical parameters, such as room shape, size, or wall materials
- *Objective*: in terms of metrics calculated from an impulse response or transfer function representative of the effect
- *Subjective*: expressing sensations such as "intimacy" or "sweetness," much like one would describe the timbre of a sound or the taste of a food.

Research on the objective characterization of the acoustics of performance spaces has led to the definition and standardization of a set of objective parameters widely adopted by acousticians, including (see e.g. [36], [40]–[43]): Reverberation decay time (Rt), Early decay time (Edt), Early-to-late energy ratio (C50, C80), Strength (G), Lateral energy fraction (LF), and Interaural cross-correlation coefficient (IACC).

Several investigations have sought to exhibit a minimal set of subjective parameters suitable for explaining or predicting listener sensations or preference judgments, or to reveal correlations between objective metrics and subjective attributes (e.g. [44]–[49]). In [41], consolidating the results of multiple studies carried out in the 1950s-80s, a concise set of subjective parameters is proposed to characterize the music concert audience experience: Loudness, Reverberance, Running Liveness/Reverberance, Clarity, Intimacy, Warmth, and Spaciousness (later split into Apparent Source Width (ASW) and Listener Envelopment (LEV) [36]);

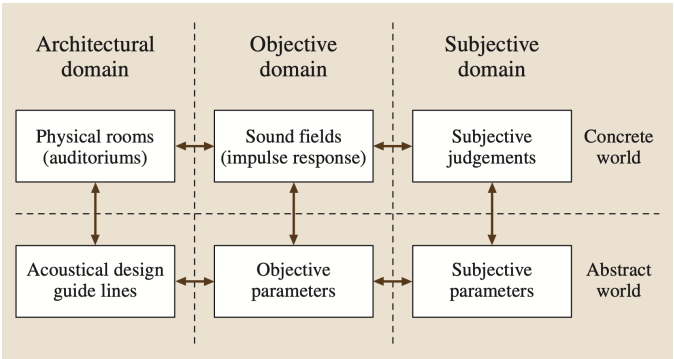

Fig. 4. Overview of concepts related to subjective room acoustics, from [36].

B. Towards a structured perceptual model for controlling virtual acoustic systems

The definition of a structured description model is a key step towards realizing a perceptually based control interface for virtual acoustic systems. Two essential properties should be fulfilled: its conciseness and its uniqueness. In such a model, a virtual acoustic condition is a point in the perceptual space described by a minimal set of mutually independent control dimensions (referred to in the following as *perceptual attributes*). This mutual independence (orthogonality) is important for the uniqueness of the description and, in practice, to ensure that modifying the setting of a given perceptual attribute will affect as little as possible the sensations controlled by adjusting other perceptual attributes in the set.

These objectives motivated a series of perceptual tests conducted in the late 1980s by IRCAM [50]–[53]. This extensive psycho-experimental research initiative ultimately led to the design of the perceptual control interface module implemented in the Spat library, named Spat_OPer (see Fig. 5 and Table I). Each test consisted in collecting perceptual dissimilarity judgments between virtual acoustic conditions presented in pairs. Such a non-verbal approach was preferred over pre-established questionnaires, in order to avoid possible semantic bias. From the dissimilarity judgements, the INDSCAL multidimensional analysis method was used to build a representation space whose dimensions reflect not only the variance expressed in the collected data but also the individual sensitivity of each participant along each dimension [54].

The resulting representation space allows projecting the relative positions of the different objects under study (here the different virtual acoustic conditions) and quantifying the weight that each participant assigns to each dimension. In other words, the representation space of a given participant (that would reflect his/her individual dissimilarity judgments) may be obtained by stretching or compressing the common space along its principal dimensions.

TABLE I: The set of nine independent acoustical criteria (left) and corresponding perceptual factors (right) implemented in the Spat_OPer control interface (see Fig. 5). The first three represent the perception of the source itself; the next three characterize the interaction between the source and the room; the last three control the room's late reverberation decay.

Acoust. criterion	Range	Unit	Perceptual factor	Range	Sensitivity
Es	[-40 0]	dB	Source Presence	[0 120]	4/dB
Desl	[-10 10]	dB	Source Warmth	[0 60]	3/dB
Desh	[-10 10]	dB	Source Brilliance	[0 60]	3/dB
Rev	[-40 0]	dB	Room Presence	[0 120]	3/dB
Edt	slave	s	Running Reverb.	[0 50]	slave
Rd1	slave	dB	Envelopment	[0 50]	slave
Rt	[0.1 10]	s	Reverberance	[0 100]	5/dBs
Drtl	[0.1 10]	–	Heaviness	[0 50]	2.5/dB
Drth	[0.1 1]	–	Liveness	[0 50]	5/dB

C. Overview of psycho-experimental procedure and results

In order to bound test duration and complexity, each perceptual experiment was based on a limited set of acoustical conditions (7 to 10), thus only spanning a small portion of the whole perceptual space. In total, sixteen tests were conducted [50]. Virtual acoustic conditions were varied along different classical architectural acoustic criteria (e.g. Rt, C80, G, ...) or low-level parameters (time and spatial distribution of the first and late reflections). They were applied to musical excerpts (solo instrument or singer) and were synthesized with a set of digital audio processors available at the time (delay units, gains, filters, reverberator). Tests concentrating on time and spectral distribution were generally conducted on headphones, whereas tests involving the manipulation of the spatial distribution of synthetic reflections were conducted in IRCAM's anechoic room equipped with a dome of nine loudspeakers (seven in the horizontal plane and two at elevated positions in the median plane [50]).

The analysis of each of the sixteen tests provided a partial representation space driven by two to four axes, yielding a total of about 50 dimensions. The next step of the study was to classify these dimensions, trying both to detect which were common to several tests and to determine their optimal objective correlation over the full set of tests [52]. The analysis led to the definition of ten perceptual dimensions whose objective definitions are reported in Appendix A and whose labels were proposed by a small group of expert listeners. Of these ten dimensions, nine were retained in the practical implementation of the Spat processor (details in Appendix B). It should be noted that the expression of the objective quantities that drive the perceptual attributes are non-trivial. Some of them include conditional contributions of time segment energy (such as described in [55]), energy masking effects and spatial dependence of late reflections. Masking effects proved critical, for instance, to formulate the loudness of the reverberation (or *Room Presence*) and to differentiate *Running Reverberance* (reverberance audible while music is playing) from (*Late*) *Reverberance* (reverberance audible after music stops) – see Table I and [56]–[58].

As observed in [59], a feature of the resulting perceptual model is an implicit "stream segregation" within the perception of an audio object, whereby the early and late energies that shape the impulse response are associated with two independent percepts (we note that this emergent property is analogous to the subjective parsing of sound events into *foreground* and *background* components observed in [60]). The perceptual factors Source Presence and Room Presence (see Table I and Fig. 5) correlate to the principal sensations that vary depending on the position or directivity of a sound source within a given room. In the development and applications of the Spat processor, reviewed in the next section, it was observed in particular that isolated control of the perceptual factor Source Presence emulates with remarkable effectiveness a variation in source-to-listener distance.

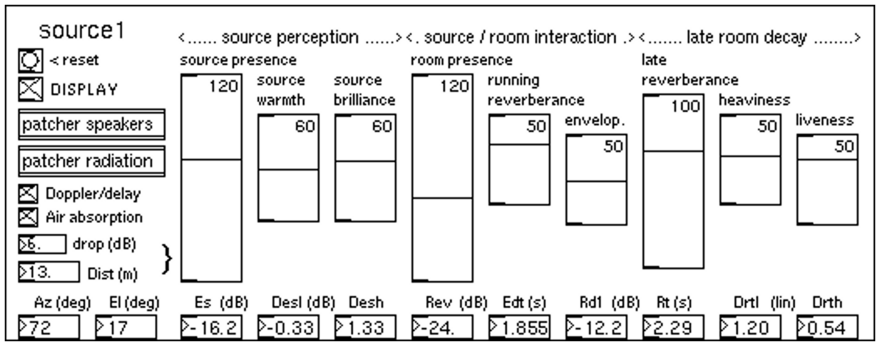

Fig. 5. Original Spat_OPer high-level perceptual control interface for a single sound source [61].

IV. The Spat Audio Processing Suite

At its inception, the *Spatialisateur* project aimed to develop a spatial reverberation processing system controlled directly via the set of perceptual attributes uncovered by the psycho-experimental study reported in section III – which yields an efficient high-level parametrization of the perception of room acoustical effects in concert halls, opera houses or auditoria. In this section, we review the original design and subsequent enhancements of the Spat library developed by IRCAM in the Max/MSP programming environment [29], [62].

A. Derivation of low-level reverberation rendering parameters

Each one of the acoustical criteria which together form the high-level description of the reverberation effect can be expressed from energetic measures derived from a decomposition of the impulse response over three frequency bands and four temporal sections (Fig. 6), assuming:

- frontal incidence of the direct sound (R0)
- random incidence of the early reflections (R1) within a frontal planar sector (60-degrees wide)
- isotropic diffuse distribution of late reflections (R2) and reverberation tail (R3)
- time limits (l1, l2, l3) equal to 20, 40 and 100 ms relative to the time of arrival of the direct sound
- reference low and high frequencies (fl, fh) equal respectively to 250 Hz and 4 kHz.

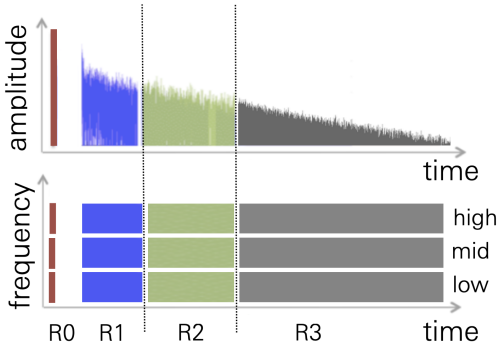

Fig. 6. Time-frequency-energy decomposition of renderer response (time limits: [20, 40, 100 ms]; reference frequencies: [250, 4000 Hz]).

Appendix A reviews the set of equations that relate the high-level acoustical criteria to these energetic measures, per [53]. The implementation of the perceptual control paradigm requires performing the inverse conversion in order to recompute the time-frequency-energy distribution following an update in the settings of the acoustical criteria. This high-level to low-level control mapping function executes the calculations reviewed in Appendix B [15], [63]. Each acoustical criterion is paired with a perceptual factor exposed in the Spat_OPer control panel. Based on this nonlinear mapping and the variance associated with each dimension returned by the INDSCAL analysis (section III-B), the relative sensitivity of the perceptual factors was averaged across test participants. This is accounted for in the scaling of each slider in the graphical user interface (see Fig. 5 and Table I).

Optionally, the set of calculated low-level processing parameters values may be further corrected by a *context compensation* operator (Fig. 7) which automatically performs, in the time-frequency-energy domain, a deconvolution accounting for the acoustic response of a loudspeaker reproduction environment [63], [64]. This compensation is based on principles similar to those underlying the object-based "room-in-room" acoustic response correction method developed in [65], [66].

B. Spat renderer algorithm for a single sound source

In addition to room acoustical descriptors, the high-level user interface Spat_OPer exposes controls for the localization and the radiation of the sound source, as shown in Fig. 7 [13], [61]. The internal topology of the core rendering functions, the *Room* and *Panning* modules, is shown in Fig. 9 [13], [64], [67]. An example of impulse response produced by the system is displayed in Fig. 8.

The four elementary filters (R0...R3) are 3-band spectral correctors [68] that shape the time-frequency-energy distribution in the reverberator response following adjustments of the perceptual factors. Additionally, for a directional sound source, their gain settings are offset according to the source's radiation parameters: its orientation and its frequency-dependent directivity pattern, defined by its omni and axis spectral corrections and an *"inside cone"* aperture angle parameter (for details, see [1] about parameter definition and Appendix B about implementation in Spat).

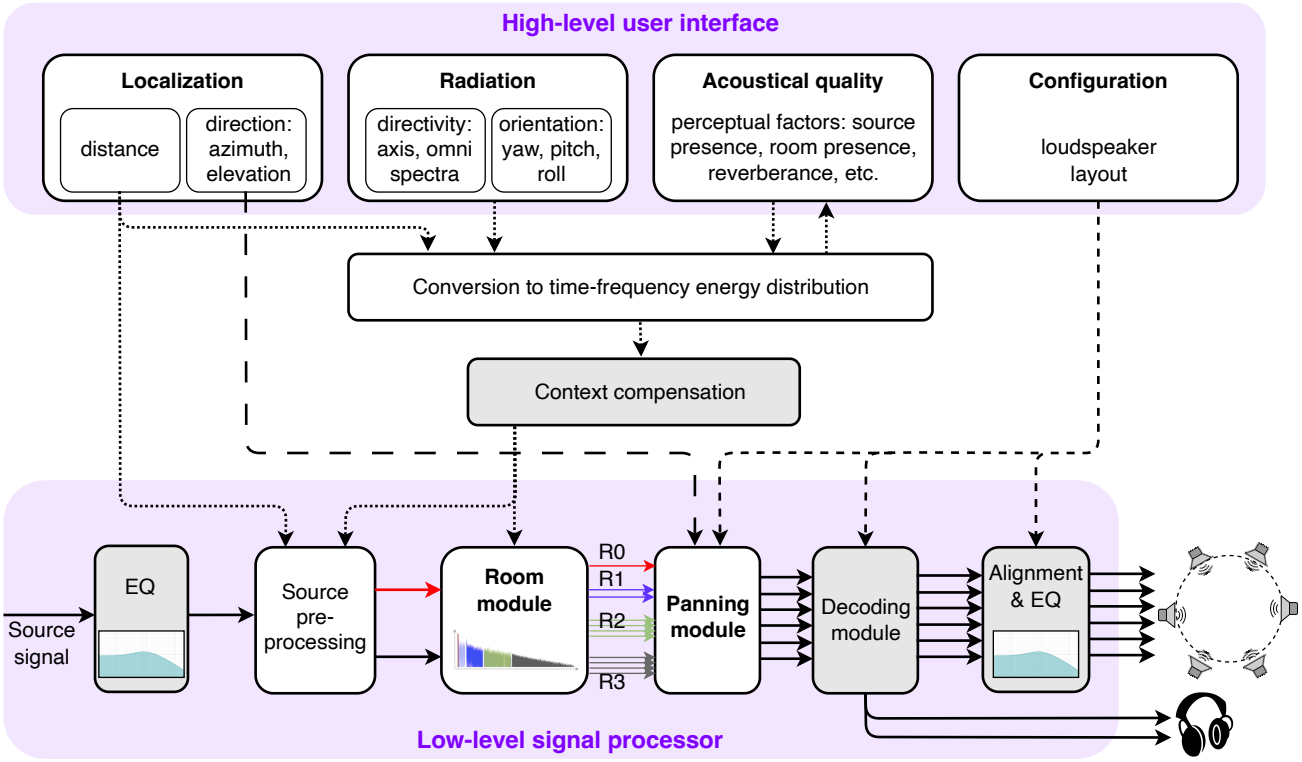

Fig. 7. Structure of a single-source Spat processing module, showing the control paths from application-level representation to low-level renderer parameters. Modules in gray color are optional.

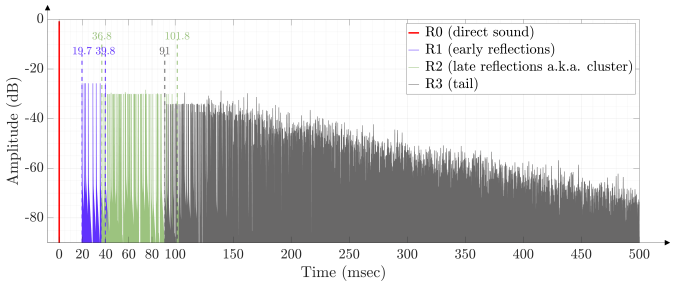

Fig. 8. Spat impulse response for a particular setting of the low-level temporal settings min, max and distr of the early, cluster and reverb segments (see user control interface in Fig. 10(d)).

The *Panning* module controls the perceived direction of sound arrival relative to the listener's head orientation, by adapting the output signals of the *Room* module to a chosen positional audio rendering method. In its initial release, the Spat library supported several panning techniques (including stereo, pairwise, ambisonic and binaural) and variable planar loudspeaker configurations [61], [69]. A nonzero setting of the azimuth angle is equivalent to a rotation of the listener's head in the opposite direction, applied to both the direct sound (R0) and the synthetic early reflections (R1) by the elementary *panner* modules (see Fig. 9). This preserves the spatial relations between the components (R0...R3) per section IV-A, thus ensuring the conservation of the acoustical criterion Es and of the perceptual factor Source Presence. Independently,

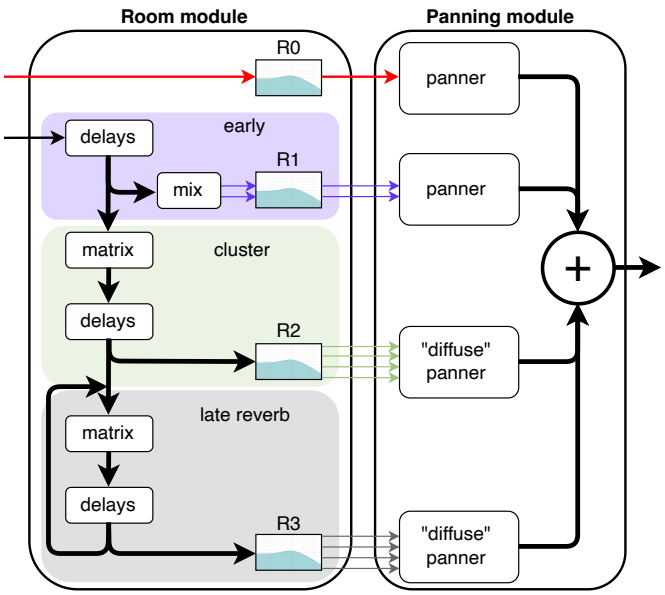

Fig. 9. Core Spat rendering algorithm architecture (*Room* and *Panning* modules) for a single sound source. Thicker lines indicate multichannel audio signal paths.

as noted at the end of section III-C, the perceived source-to-listener distance is controlled by offsetting Es so that increasing distance has the effect of reducing the energies R0 and R1, per equations (12), (13) of Appendix A.

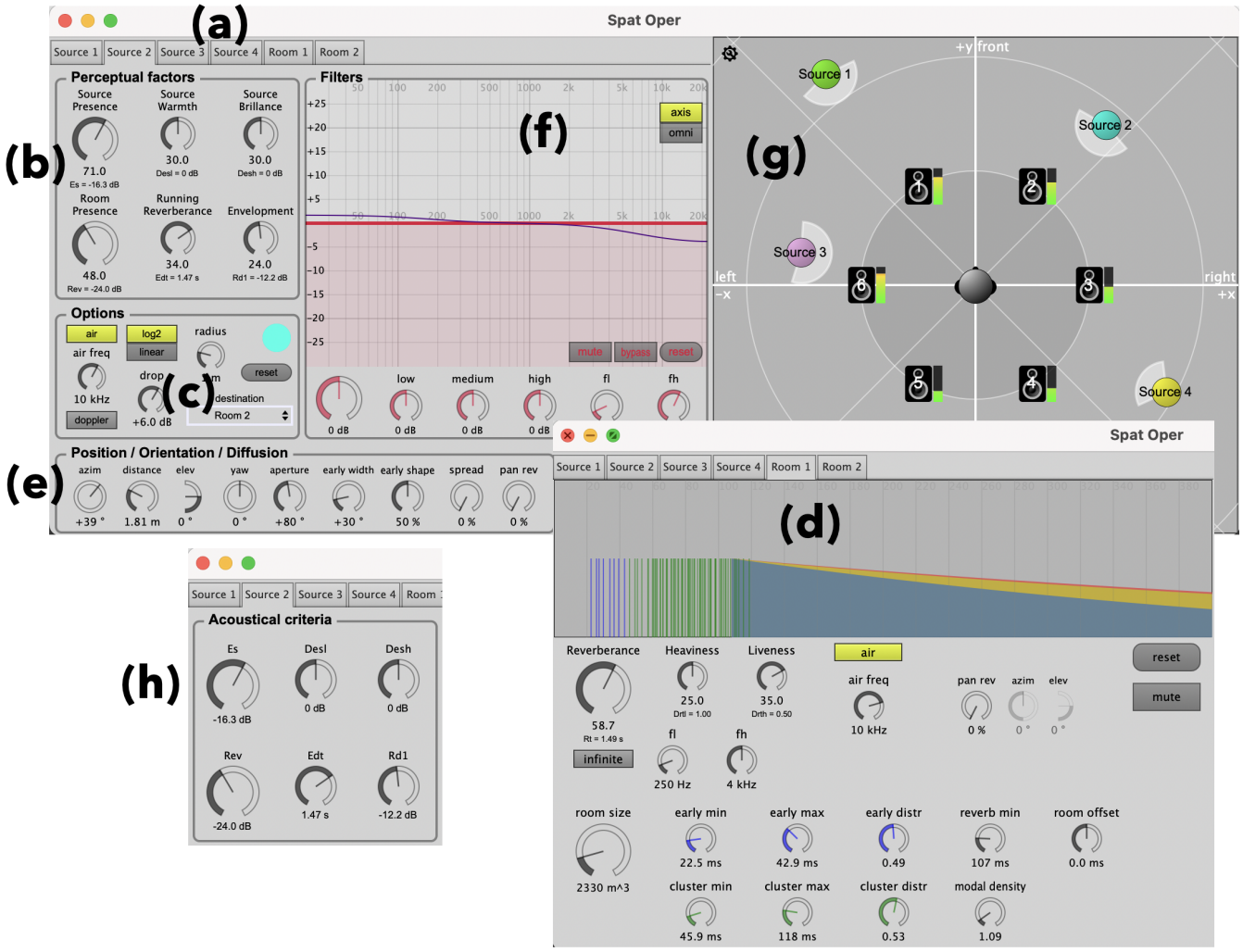

Fig. 10. Multi-object, multi-room Spat_OPer control panel. (a) Tabs for selecting sources and rooms; (b) Perceptual factors for a selected source; (c) Assignment of a source to a room; (d) Settings of a selected room; (e) Additional source parameters; (f) Axis and omni spectral settings for a selected source; (g) Graphical representation of the scene (top-down view); (h) Detailed view of the acoustical criteria for a selected source.

C. Extensions and evolution

Over the past decades, the Spat rendering framework and its Spat_OPer control interface were implemented in various forms – including *external* objects in the Max/MSP environment [62], [29], [70], [71]; modules in the *Open Music* and *o7* computer-aided composition frameworks [72], [73]; digital audio workstation (DAW) plugins [74]; devices for Ableton Live [75]–[77]; standalone applications [74], [78]; or hardware processor [79]. While the original design architecture (Fig. 7) was retained, several key features were extended or altered.

a) Multi-object, multi-room implementations: With the increase in available computing power, the system was extended to render multiple sound objects and several rooms concurrently. The Spat_OPer user interface was revised and extended with tabs presenting control panels for the different sources and rooms (Fig. 10). Similar capability was implemented in *Panoramix*, a standalone software application for immersive audio mixing and production (Fig. 11) [78]. When several sources are assigned to the same room, their contributions sum at the multichannel input stage of the late reverberation computation module (gray box in Fig. 9), thereby sharing its processing cost.

b) Extension from 2-D to 3-D immersive audio: The *Panning* module supports recently developed multi-channel positional audio rendering techniques (see e.g. [8], [35]), including binaural, VBAP, wave-field synthesis and high-order Ambisonics, enabling object-based immersive audio spatialization along both azimuth and elevation. This extension to 3-D loudspeaker geometries required overcoming limitations of the original design, such as the directional distribution of the early reflections (R1) and the number of audio channels needed for reproducing the spatial impression of a diffuse reverberant sound field [29]. The latter is a topic of ongoing research [80]–[83] both for channel-based reproduction (where the *"diffuse" panner* module is a diffusion matrix) and scene-based rendering (where it ambisonically encodes the R2 and R3 output channels to simulate impinging plane waves).

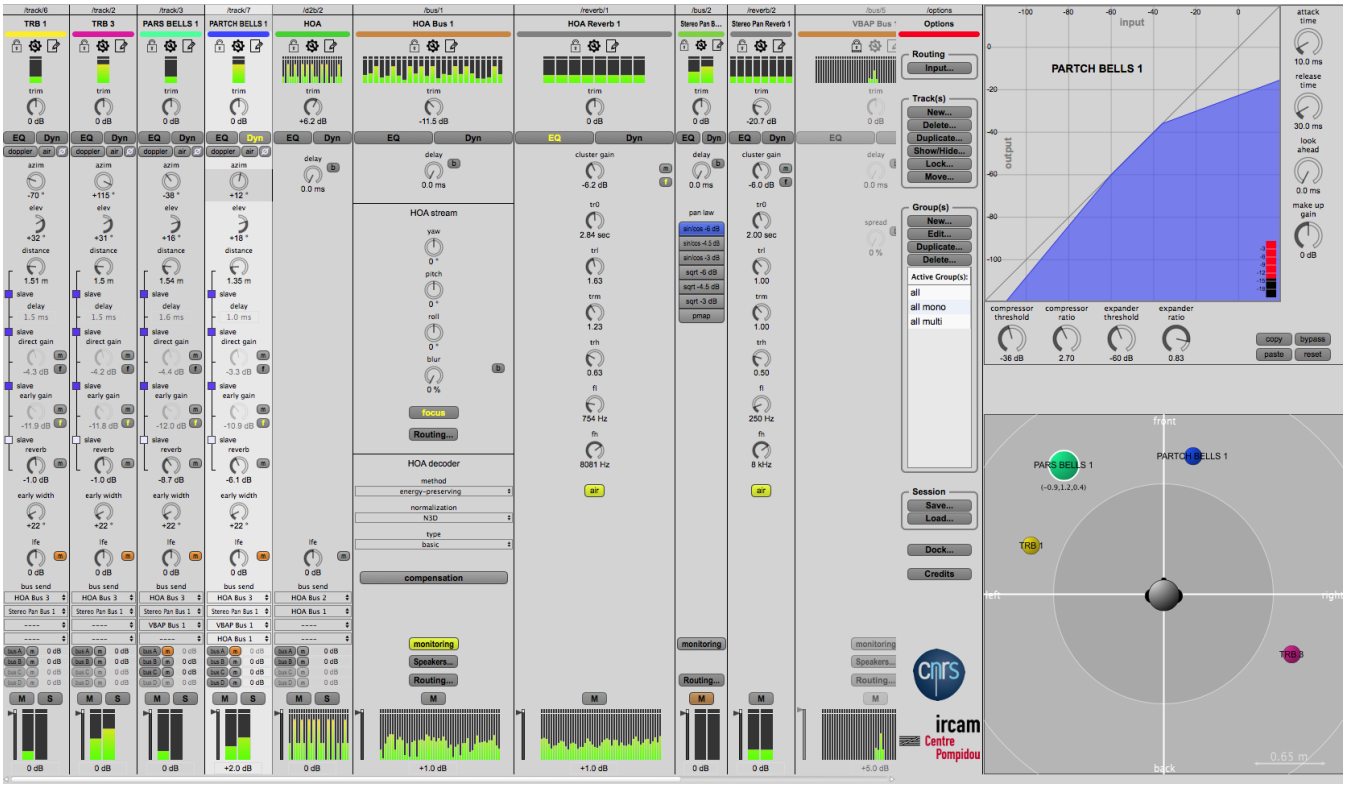

Fig. 11. Control panel of the standalone application *Panoramix* for linear audio production in various multichannel immersive formats, which supports combining 6-DoF spatial audio objects with microphone-captured or pre-recorded stereo, pairwise-panned or Ambisonic multichannel audio input tracks [78].

c) Beyond auditoria and concert halls: The psycho-experimental study presented in section III-C spanned a limited range of acoustical conditions typically representative of music stimuli in mid-size auditoria. For artistic or simulation applications, it is desirable to cover a larger class of room types (from bathrooms to cathedrals). The meta-parameter room size has the effect of stretching the time limits of the R1, R2 and R3 sections (see Figs. 8 and 10(d)) according to the desired room volume. This simple approach has proven satisfactory for some artistic usages. However, the expressions of the acoustical criteria per Eq. (2)–(11) of Appendix A are less perceptually valid when values of the time limits (l1, l2, l3) differ significantly from their default settings.

d) Hybrid reverberation and analysis/synthesis: In [84], a procedure was proposed for the time-frequency analysis/synthesis and denoising of room impulse responses, facilitating the estimation of the time-frequency-energy distribution and reverberation decay parameters that drive the Spat reverberator. By applying Equations (2)–(11) provided in Appendix A, the high-level perceptual control parameters exposed in Spat_OPer are calculated automatically to simulate an existing reverberation response. This mapping to the perceptual domain enables natural-sounding alterations and interpolations between room acoustical conditions. In order to support the authentic simulation of a measured acoustic response while exposing a highly flexible perceptually based user control, an alternative realization of the *Room* module was developed [85], in which the early response (R1, R2) is rendered by a convolution-based algorithm, whereas the time-frequency envelope of the reverberation decay (R3) is accurately reproduced by a feedback delay network (FDN) if the measured reverberation response presents an exponentially decaying time-frequency envelope [84], [86].

e) Anisotropic late reflections and reverberation: The canonical reverberation response model presented in Section IV-A assumes that late reflections (R2) and reverberation (R3) conform to an ideal isotropic sound field model agnostic to the listener's orientation. This assumption, theoretically valid inside a *mixing room* [87], does not generalize to semi-open spaces, environments presenting elongated or irregular shapes, or enclosures comprising acoustically coupled volumes [39]. Recent studies investigate the perception of the anisotropy of a reverberant field [88]–[90], objective metrics for the characterization of directional reverberation [91]–[94], compact microphone arrays for its measurement and analysis [95]–[99], and techniques for its synthesis and reproduction [99]–[102]. In Spat, the *"diffuse" panner* module (Fig. 9) implements a *divergence panning* algorithm that assigns uncorrelated input signals to an array of notional sources distributed along an arc having adjustable extent, similar to the method described in [17], [32] (see Fig. 12). Future work includes extending the automatic hybrid reverberation analysis/synthesis procedure described above to incorporate the modeling of anisotropic reverberation decays.

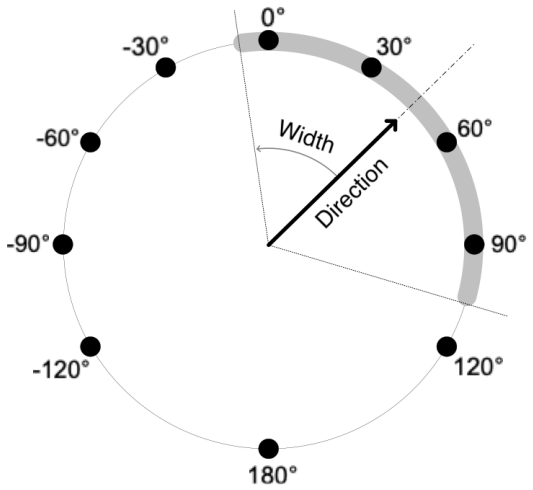

Fig. 12. Representation of the angular position and extent of a sound, employing a normalized *divergence panning vector* whose magnitude spans from 0.0 for a diffuse sound component to 1.0 for a point source [17], [32].

V. Spatial Audio Objects – Production, Coding, and Interoperability

In [1], the generic spatial audio renderer model reviewed in section II was applied to physically-based virtual world simulation scenarios such as gaming or AR/VR. In this section, we tackle its integration in a "Spatial Audio Object Workstation" realizing Spat's perceptually-based sound design functionality within a familiar audio creation platform topology and workflow [103]. We propose a metadata specification for spatial audio objects, aiming to facilitate a future unified audio content production, coding format, and experience ecosystem.

A. Perceptually-based spatial audio scene rendering

Fig. 3 illustrates a Spat-compatible configuration of the renderer, where reflections and reverberation processing resources are shared between audio objects assigned to the same room (with shared parameter controls displayed in Fig. 10(d), while the other Spat_OPer parameters remain adjustable separately for each source). This configuration exercises the following renderer properties (referring to definitions in [1]):

- Per-source processing parameters:
  - $Direct\_pan$ (azimuth and elevation angles)
  - $Reflections\_pan$ and $Reflections\_focus$
  - Effect send delays, gains, and filters (R0...R3).
- Global processing parameters
  - Effect type loaded in each of the active *effect slots*
  - Reference frequencies (Fl = 250 Hz, Fh = 4 kHz)
  - Reflections time span (for R1: 20 ms; for R2: 60 ms)
  - Reverberation $Decay\_time$, $Decay\_hf\_ratio$, $Decay\_lf\_ratio$ and $Density$ (for R3).

In order to support a perceptually-based representation, the low-level rendering interface must be listener-centric: object position and orientation coordinates are defined relative to the listener's coordinates and head pose. Although the notion of "room" is exposed at the application level (Spat_OPer interface), the renderer sees its task merely as a superposition of sound components (direct sound, reflections, reverberation), each specified via a minimal set of rendering commands. Similarly, the source's orientation and directivity parameters are not required in the low-level rendering interface.

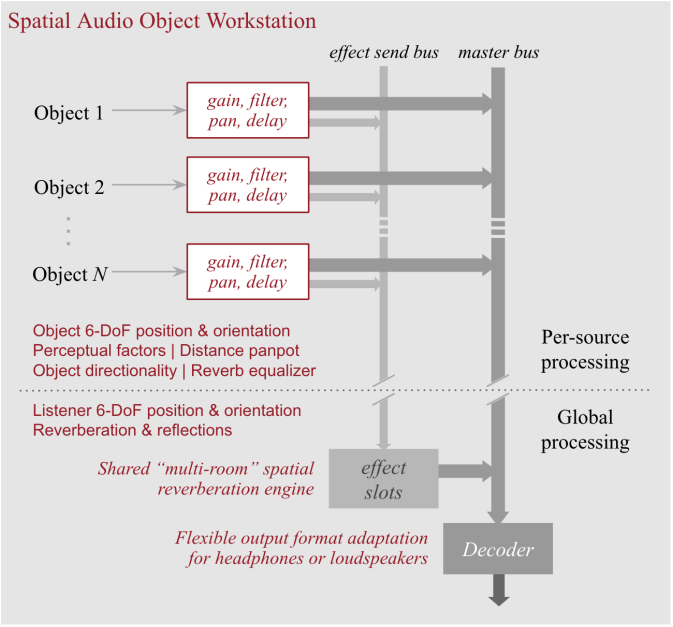

Fig. 13. Implementing a "Spatial Audio Object Workstation" by mapping the generic renderer of Fig. 3 to a standard digital audio workstation (DAW) signal flow. This exposes in a DAW environment the functionality illustrated in Figs. 10–11, including specific sound design features reviewed in section V-B. The per-object processing functions and the signal formats of the internal multi-channel summing buses, *effect send bus* and *master bus*, are apparent in Fig. 3 (incorporating in the *Decoder* module the multi-channel decorrelation filter bank that enables the rendering of spatially extended sounds).

B. Spatial Audio Object Workstation

As illustrated in Fig. 13, the renderer of Fig. 3 shares the familiar signal flow of a digital audio workstation (DAW), but features a non-standard *effect send bus* format determined by the configuration of the *effect slots*. This enhancement enables all of the features described in section IV (Figs. 10–11), including useful per-source sound design functions not commonly offered in audio production tools:

- A practical and effective *"distance panpot"* [104], realized by simply mapping to the perceptual factor Source Presence (per sections IV-B and III-C).
- Per-object control of the perceptual attributes {Source Presence, Brilliance, Warmth} and {Room Presence, Running Reverberance, Envelopment}.
- The axis spectral correction, which affects the direct-sound frequency response according to the source's orientation and to its other Radiation parameters, including:
- The omni spectral correction, which, importantly, enables fine-tuning the spectral contour of the reverberation and reflections individually for each sound source (simulating its *diffuse-field transfer function* [1], [33], [84]).

C. Spatial audio objects: interoperability requirements

Within the ecosystem represented in Fig. 2, 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 an application independent of the creation system. The *scene description* function converts the application-specific scene representation to a vendor-neutral specification compatible with the *encoder* or the 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 rendering engine. For each 6-DoF 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 VR/AR, 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 strategies are possible:

- *Object-based:* incorporating reverberation processing parameters within the object's metadata, as in:
  - *MPEG-4 Advanced AudioBIFS (Perceptual Approach)*, with a parametrization based on Spat [14], [15]
  - *RSAO (Reverberant Spatial Audio Object)* [18], which provides a set of discrete reflections combined with a parametric reverberation decay tail.
- *Environment-based:* in this case, a separate metadata interface describes the acoustic environment, as in:
  - *MPEG-4 Advanced AudioBIFS (Physical Approach)*, which is based on the DIVA research project [3], [14]
  - *OpenAL Effects Extension*, an audio programming interface providing immersive reverberation as a shared parametric effect processing option [1], [16], [31]
  - *MPEG-I*, which includes the constructs *Acoustic Environment* and *Geometry*, separate from *Sources* [19].

In the second case, 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. [1], [3], [5]).

D. 6-DoF spatial audio objects: a proposal

In defining a generic 6-DoF spatial audio scene rendering interface, our objectives are: (a) allowing maximum flexibility in per-object sound design by the content creator; (b) ensuring faithful reproduction of spatial audio scenes at the playback end; (c) letting the end-user take advantage of the interactivity and personalization benefits enabled by the object-based representation, within limits authorized by the content creator; (d) enabling both the content creator and the playback application developer to select among several possible levels of end-user interaction and implementation complexity.

In Table II, we categorize the properties of a spatial audio object into a hierarchy of *functionality profiles*, exposing increasing levels of playback-end spatial audio scene interaction.

TABLE II: Proposed hierarchy of metadata functionality profiles in a generic 6-DoF spatial audio object encoding specification.

"Basic"	Minimum object-based metadata set required to ensure faithful rendering of the transmitted spatial audio scene (also optionally allows gain scaling, removal, substitution, or 3-DoF repositioning of any individual object):
	Position (azim, elev) and *Extent* of the sound source may be jointly represented by a normalized *divergence panning vector* (p0) as in [17], [32] (see section IV-C(e) and Fig. 12);
	Reverberation response: time-frequency-energy distribution (or acoustical criteria, per Appendix A), time limits (l1, l2, l3), and divergence panning vectors (p1, p2, p3).
"Remix"	Additional object-based metadata to enable and control spectral equalization (useful e.g. in playback-end remixing, mashup or karaoke application scenarios):
	Spectral corrections: axis (correcting the object's direct sound) and omni (correcting its reverberation and reflections).
"Navig."	Additional object-based metadata to enable and control playback-end 6-DoF navigation and repositioning of individual objects (using e.g. Spat_OPer) [1], [31]:
	Orientation coordinates and off-axis directivity pattern;
	Distance in "world units," and optional parameters for controlling *distance-based attenuation factors* (see e.g. [1], [31]);
	Optional parameters for controlling features such as *air absorption*, *Doppler effect* or *propagation delay*.
"Envir."	Separate *environment-based* metadata (see section V-C) describing one or more acoustic spaces and enabling geometrical manipulations of the audio scene (e.g. under control of a virtual world representation [20]), with the option to override or omit the reverberation metadata (defined above) for some of the objects.

VI. Conclusion

In this paper, we proposed solutions towards unifying spatial audio content creation, distribution and playback platforms:

- *Rendering:* we verified that the generic interface described in [1] is compatible with IRCAM Spat, taken as an example of a perceptually-based immersive music creation tool.
- *Production*: the "Spatial Audio Object Workstation" maps to the familiar DAW signal flow and extends it with a useful distance (or "depth") control enabling 6-DoF positioning, with or without an underlying physical environment model.
- *Coding*: we propose a 6-DoF spatial audio object specification ensuring faithful scene reproduction in flexible playback configurations, while supporting enhanced consumer-end interaction including remixing, navigation, personalization or compensation of listening conditions or acoustics.

Future directions include: realizing proof-of-concept AR/VR and networked music performances demonstrating the above opportunities, exploring the combination of physical and perceptual application-level scene representations, and evaluating the compatibility of the above proposals with extended or alternative perceptual models of room reverberation.

VII. Acknowledgments

The authors are grateful to Earl Vickers for his thoughtful suggestions, which helped improve the clarity of this paper.

References

[1] J.-M. Jot, R. Audfray, M. Hertensteiner, and B. Schmidt, "Rendering Spatial Sound for Interoperable Experiences in the Audio Metaverse," in *Proc. Int. Conf. Immersive and 3D Audio (I3DA)*, Bologna, Italy, September 2021. https://doi.org/10.1109/I3DA48870.2021.9610971

[2] M. Kleiner, B.-I. Dalenbäck, and P. Svensson, "Auralization – An Overview," *J. Audio Eng. Soc.*, vol. 41, no. 11, pp. 861 – 875, November 1993. www.aes.org/e-lib/browse.cfm?elib=6976

[3] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen, "Creating Interactive Virtual Acoustic Environments," *J. Audio Eng. Soc.*, Sep 1999. www.aes.org/e-lib/browse.cfm?elib=12095

[4] M. Vorländer, *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. Berlin: Springer, 2008. https://doi.org/10.1007/978-3-030-51202-6

[5] N. Raghuvanshi and J. Snyder, "Parametric Directional Coding for Precomputed Sound Propagation," *ACM Trans. Graphics*, vol. 37, no. 4, September 2018. https://doi.org/10.1145/3197517.3201339

[6] F. Pind, J. Einarsson, S. Guojonsson, M. Cosnefroy, J. Pedersen, J. Stefansson, and A. Milo, "A Novel Wave-Based Virtual Acoustics and Spatial Audio Framework," in *Proc. AES Conf. Audio for Virtual and Augmented Reality*, Seattle, CA, USA, August 2022. www.aes.org/e-lib/browse.cfm?elib=21871

[7] K. Godin, H. Gamper, and N. Raghuvanshi, "Aesthetic Modification of Room Impulse Responses for Interactive Auralization," in *Proc. AES Int. Conf. Immersive and Interactive Audio*, York, UK, March 2019. www.aes.org/e-lib/browse.cfm?elib=20444

[8] A. Roginska and P. Geluso, *Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio*. Focal Press, 2017. https://doi.org/10.4324/9781315707525

[9] J. Wyner, *Audio Mastering - Essential Practices*. Berklee Press, 2013.

[10] C. Q. Robinson, S. Mehta, and N. Tsingos, "Scalable Format and Tools to Extend the Possibilities of Cinema Audio," *SMPTE Motion Imaging Journal*, vol. 121, no. 8, pp. 63–69, 2012. https://doi.org/10.5594/j18248XY

[11] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H Audio–The New Standard for Universal Spatial/3D Audio Coding," in *Proc. 137th Convention of the Audio Eng. Soc. (AES)*, Los Angeles, CA, USA, October 2014. www.aes.org/e-lib/browse.cfm?elib=17556

[12] J.-M. Jot and Z. Fejzo, "Beyond Surround Sound - Creation, Coding and Reproduction of 3-D Audio Soundtracks," in *Proc. 131st Convention of the Audio Eng. Soc. (AES)*, October 2011. www.aes.org/e-lib/browse.cfm?elib=15989

[13] J.-M. Jot, "Real-Time Spatial Processing of Sounds for Music, Multimedia and Interactive Human-Computer Interfaces," *ACM Multimedia Systems Journal*, vol. 7, no. 1, pp. 55 – 69, January 1999. https://doi.org/10.1007/s005300050111

[14] R. Väänänen and J. Huopaniemi, "Advanced AudioBIFS: virtual acoustics modeling in MPEG-4 scene description," *IEEE Transactions on Multimedia*, vol. 6, no. 5, pp. 661 – 675, September 2004. https://doi.org/10.1109/TMM.2004.834864

[15] ISO (Int. Org. for Standardization), "ISO/IEC 14496-11:2005: MPEG-4 – Part 11: Scene description and application engine," December 2005. www.iso.org/standard/38560.html

[16] J.-M. Trivi and J.-M. Jot, "Rendering MPEG-4 AABIFS content through a low-level cross-platform 3D audio API," in *Proc. IEEE Int. Conf. Multimedia and Expo*, vol. 1, Lausanne, Switzerland, August 2002, pp. 513 – 516. https://doi.org/10.1109/ICME.2002.1035831

[17] J.-M. Jot, "Interactive 3D audio rendering in flexible playback configurations," in *Proc. Asia Pacific Signal and Information Processing Association Annual Summit and Conf.*, December 2012. http://www.apsipa.org/proceedings_2012/papers/401.pdf

[18] P. Coleman, A. Franck, P. J. Jackson, R. J. Hudges, L. Remaggi, and F. Melchior, "Object-based reverberation for spatial audio," *J. Audio Eng. Soc.*, vol. 65, no. 1/2, pp. 66 – 77, January/February 2017. https://doi.org/10.17743/jaes.2016.0059

[19] J. Herre and S. Disch, "MPEG-I Immersive Audio – Reference Model For The Virtual/Augmented Reality Audio Standard," *J. Audio Eng. Soc.*, vol. 71, no. 5, pp. 229 – 240, May 2023. https://doi.org/10.17743/jaes.2022.0074

[20] L. Turchet, "Musical Metaverse: vision, opportunities, and challenges," *Personal and Ubiquitous Computing*, January 2023. https://doi.org/10.1007/s00779-023-01708-1

[21] N. Meyer-Kahlen, P. Piiroinen, G. Vishwanath, P. Juntunen, E. Tiainen, and S. J. Schlecht, "Inside The Quartet - A first-person virtual reality string quartet production," in *Proc. 154th Convention of the Audio Eng. Soc. (AES)*, Espoo, Finland, May 2023. www.aes.org/e-lib/browse.cfm?elib=22078

[22] P. Cairns, A. Hunt, D. Johnston, J. Cooper, B. Lee, H. Daffern, and G. Kearney, "Evaluation of Metaverse Music Performance With BBC Maida Vale Recording Studios," *J. Audio Eng. Soc.*, vol. 71, no. 6, pp. 313 – 325, June 2023. https://doi.org/10.17743/jaes.2022.0086

[23] L. Turchet, R. Hamilton, and A. Çamci, "Music in Extended Realities," *IEEE Access*, vol. 9, pp. 15 810 – 15 832, January 2021. https://doi.org/10.1109/ACCESS.2021.3052931

[24] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An Overview on Networked Music Performance Technologies," *IEEE Access*, vol. 4, pp. 8823 – 8843, 2016. https://doi.org/10.1109/ACCESS.2016.2628440

[25] B. Loveridge, "Networked Music Performance in Virtual Reality: Current Perspectives," *J. Network Music and Arts*, vol. 2, no. 1, pp. 1 – 19, 2020. https://commons.library.stonybrook.edu/jonma/vol2/iss1/2/

[26] R. Hupke, S. Preihs, and J. Peissig, "Immersive Room Extension Environment for Networked Music Performance," in *Proc. 153rd Convention of the Audio Eng. Soc. (AES)*, Online, October 2022. www.aes.org/e-lib/browse.cfm?elib=21909

[27] A. Genovese, "Acoustics and Copresence: towards effective auditory virtual environments for distributed music performances," Ph.D. dissertation, New York University, New York City, USA, 2023.

[28] K. Parkkola, T. McKenzie, J. Häkkinen, and V. Pulkki, "Representing Inner Voices in Virtual Reality Environments," in *Proc. 154th Convention of the Audio Eng. Soc. (AES)*, Espoo, Finland, May 2023. www.aes.org/e-lib/browse.cfm?elib=22046

[29] T. Carpentier, M. Noisternig, and O. Warusfel, "Twenty Years of Ircam Spat: Looking Back, Looking Forward," in *Proc. 41st Int. Computer Music Conference (ICMC)*, Denton, TX, USA, September 2015, pp. 270 – 277. http://hdl.handle.net/2027/spo.bbp2372.2015.056

[30] J.-M. Jot and J.-M. Trivi, "Scene Description Model and Rendering Engine for Interactive Virtual Acoustics," in *Proc. 120th Convention of the Audio Eng. Soc. (AES)*, Paris, France, May 2006. www.aes.org/e-lib/browse.cfm?elib=13464

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

[32] J.-M. Jot, A. Philp, and M. Walsh, "Binaural Simulation of Complex Acoustic Scenes for Interactive Audio," in *Proc. 121st Convention of the Audio Eng. Soc. (AES)*, San Francisco, CA, USA, October 2006. www.aes.org/e-lib/browse.cfm?elib=13784

[33] J.-M. Jot and K.-S. Lee, "Augmented Reality Headphone Environment Rendering," in *Proc. AES Conf. Audio for Virtual and Augmented Reality*, Los Angeles, CA, USA, September 2016. www.aes.org/e-lib/browse.cfm?elib=18506

[34] Z. Ben-Hur, D. L. Alon, R. Mehra, and B. Rafaely, "Binaural Reproduction Based on Bilateral Ambisonics and Ear-Aligned HRTFs," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 29, January 2021. https://doi.org/10.1109/taslp.2021.3055038

[35] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 1920 – 1938, September 2013. https://doi.org/10.1109/jproc.2013.2264784

[36] A. C. Gade, "Acoustics in Halls for Speech and Music," in *Handbook of Acoustics*, T. D. Rossing, Ed. Springer, 2007, pp. 301 – 350. https://doi.org/10.1007/978-0-387-30425-0

[37] L. L. Beranek, *Concert Halls and Opera Houses: Music, Acoustics, and Architecture*, 2nd ed. New York, NY, USA: Springer, 2004. https://doi.org/10.1007/978-0-387-21636-2

[38] M. Barron, *Auditorium Acoustics and Architectural Design*, 2nd ed. New York, NY, USA: Spon Press, 2010. https://doi.org/10.4324/9780203874226

[39] H. Kuttruff, *Room Acoustics*, 6th ed. Boca Raton, FL, USA: CRC Press, 2016. https://doi.org/10.1201/9781315372150

[40] M. Barron and L. Lee, "Energy relations in concert auditoriums. I," *J. Acoustical Society of America*, vol. 84, no. 2, pp. 618 – 628, August 1988. https://doi.org/10.1121/1.396840

[41] L. L. Beranek, "Concert hall acoustics – 1992," *J. Acoustical Society of America*, vol. 92, no. 1, pp. 1 – 39, July 1992. https://doi.org/10.1121/1.404283

[42] ISO (Int. Org. for Standardization), "ISO 3382-1:2009: Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces," June 2009. www.iso.org/standard/40979.html

[43] M. Vorländer and F. P. Mechel, "Room acoustics," in *Formulas of Acoustics*, 2nd ed., F. P. Mechel, Ed. Berlin, Germany: Springer Verlag, 2008, ch. M, pp. 873 – 943. https://doi.org/10.1007/978-3-540-76833-3

[44] M. R. Schroeder, D. Gottlob, and K. F. Siebrasse, "Comparative study of European concert halls: correlation of subjective preference with geometric and acoustic parameters," *J. Acoustical Society of America*, vol. 56, no. 4, October 1974. https://doi.org/10.1121/1.1903408

[45] M. Barron, "Using the standard on objective measures for concert auditoria, ISO 3382, to give reliable results," *Acoustical Science and Technology*, vol. 26, no. 2, pp. 162 – 169, 2005. https://doi.org/10.1250/ast.26.162

[46] T. Lokki, J. Pätynen, A. Kuusinen, and S. Tervo, "Disentangling preference ratings of concert hall acoustics using subjective sensory profiles," *J. Acoustical Society of America*, vol. 132, no. 5, pp. 3148 – 3161, November 2012. https://doi.org/10.1121/1.4756826

[47] T. Lokki, "Tasting music like wine: Sensory evaluation of concert halls," *Physics Today*, vol. 67, no. 1, pp. 27 – 32, January 2014. https://doi.org/10.1063/pt.3.2242

[48] Y. Ando, *Concert Hall Acoustics*, M. R. Schroeder, Ed. Heidelberg, Germany: Springer, 1985. https://doi.org/10.1007/978-3-642-69810-1

[49] ——, *Opera House Acoustics Based on Subjective Preference Theory*. Springer, 2015. https://doi.org/10.1007/978-4-431-55423-3

[50] C. Lavandier, "Validation perceptive d'un modèle objectif de caractérisation de la qualité acoustique des salles," Ph.D. dissertation, Université du Maine, Le Mans, France, June 1989.

[51] J.-P. Jullien, E. Kahle, S. Winsberg, and O. Warusfel, "Some Results on the Objective Characterisation of Room Acoustical Quality in both Laboratory and Real Environments," in *Proceedings of the Institute of Acoustics*, vol. 14, no. 2, 1992.

[52] J.-P. Jullien, "Structured Model for the Representation and the Control of Room Acoustical Quality," in *Proc. 15th Int. Congress on Acoustics (ICA)*, Trondheim, Norway, June 1995, pp. 517 — 520.

[53] E. Kahle, "Validation d'un modèle objectif de la perception de la qualité acoustique dans un ensemble de salles de concerts et d'opéras," Ph.D. dissertation, Laboratoire d'Acoustique de l'Université du Maine, Le Mans, France, June 1995.

[54] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition," *Psychometrika*, vol. 35, pp. 283 – 319, September 1970. https://doi.org/10.1007/BF02310791

[55] J. P. A. Lochner and J. F. Burger, "The subjective masking of short time delayed echos by their primary sound and their contribution to the intelligibility of speech," *Acustica*, vol. 8, no. 1, pp. 1 – 10, 1958.

[56] D. Griesinger, "IALF-Binaural Measures of Spatial Impression and Running Reverberance," in *Proc. 92nd Conv. Audio Eng. Soc.*, Vienna, Austria, March 1992. www.aes.org/e-lib/browse.cfm?elib=6841

[57] E. Green, E. Kahle, V. Berrier, and E. Carayol, "Beyond 80ms: The Subjective Effects of Sound Energy Arriving Shortly After the "Early" Sound Period," in *Proc. Int. Symposium on Room Acoustic (ISRA)*, Amsterdam, Netherlands, September 2019, pp. 409 – 416. https://doi.org/10.18154/RWTH-CONV-240133

[58] A. Haapaniemi and T. Lokki, "Identifying concert halls from source presence vs room presence," *J. Acoustical Society of America*, vol. 135, no. 6, pp. EL311 – EL317, May 2014. https://doi.org/10.1121/1.4879671

[59] E. Kahle, "Room Acoustical Quality of Concert Halls: Perceptual Factors and Acoustic Criteria — Return from Experience," *Building Acoustics*, vol. 20, no. 4, pp. 265 – 282, December 2013. https://doi.org/10.1260/1351-010x.20.4.265

[60] D. Griesinger, "How loud is my reverberation?" in *Proc. 98th Audio Eng. Soc. Convention*, Paris, France, February 1995. www.aes.org/e-lib/browse.cfm?elib=7823

[61] J.-M. Jot and T. Caulkins, *Spat Reference Manual*. IRCAM, 1995. https://support.ircam.fr/docs/spat/3.0/spat-3-ref/

[62] Max/MSP by Cycling'74. https://cycling74.com

[63] J.-M. Jot, J.-P. Jullien, and O. Warusfel, "Patent US5812674 – Method to simulate the acoustical quality of a room and associated audio-digital processor," August 1996.

[64] J.-M. Jot, "Efficient models for reverberation and distance rendering in computer music and virtual audio reality," in *Proc. Int. Computer Music Conference (ICMC)*, Thessaloniki, Greece, September 1997. http://hdl.handle.net/2027/spo.bbp2372.1997.064

[65] D. Menzies and F. M. Fazi, "A Perceptual Approach to Object-Based Room Correction," in *Proc. 141st Convention of the Audio Eng. Soc. (AES)*, Los Angeles, CA, USA, September 2016. www.aes.org/e-lib/browse.cfm?elib=18391

[66] D. Menzies, P. Coleman, and F. M. Fazi, "A Room Compensation Method by Modification of Reverberant Audio Objects," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 239 – 252, 2021. https://doi.org/10.1109/taslp.2020.3036781

[67] J.-M. Jot, V. Larcher, and O. Warusfel, "Digital Signal Processing Issues in the Context of Binaural and Transaural Stereophony," in *Proc. 98th Convention of the Audio Eng. Soc. (AES)*, February 1995. www.aes.org/e-lib/browse.cfm?elib=7786

[68] J.-M. Jot, "Proportional Parametric Equalizers – Application to Digital Reverberation and Environmental Audio Processing," in *Proc. 139th Convention of the Audio Eng. Soc. (AES)*, New York, NY, USA, October 2015. www.aes.org/e-lib/browse.cfm?elib=17916

[69] J.-M. Jot and O. Warusfel, "A Real-Time Spatial Sound Processor for Music and Virtual Reality Applications," in *Proc. 21st Int. Computer Music Conference (ICMC)*, Banff, Canada, September 1995, pp. 294 – 295. http://hdl.handle.net/2027/spo.bbp2372.1995.089

[70] T. Carpentier, "A new implementation of Spat in Max," in *Proc. 15th Sound and Music Computing Conference (SMC)*, Limassol, Cyprus, July 2018, pp. 184 – 191. https://doi.org/10.5281/zenodo.1422552

[71] ——, "Spat: a comprehensive toolbox for sound spatialization in Max," *Ideas Sónicas*, vol. 13, no. 24, pp. 12 – 23, June 2021. https://hal.science/hal-03356292

[72] J. Garcia, T. Carpentier, and J. Bresson, "Interactive-compositional authoring of sound spatialization," *J. New Music Research – Special Issue on Interactive Composition*, vol. 46, no. 1, pp. 74 – 86, 2017. https://doi.org/10.1080/09298215.2016.1230632

[73] J. Bresson, D. Bouche, T. Carpentier, D. Schwarz, and J. Garcia, "Next-generation Computer-aided Composition Environment: A New Implementation of OpenMusic," in *Proc. Int. Computer Music Conference (ICMC)*, Shanghai, China, October 2017. http://hdl.handle.net/2027/spo.bbp2372.2017.042

[74] Spat Revolution by Flux::. www.flux.audio/project/spat-revolution

[75] Ableton Live. www.ableton.com

[76] Spat for Live by MusicUnit. www.ableton.com/packs/spat-bundle

[77] XP4L. www.xp4l.com

[78] T. Carpentier, "Panoramix: 3D mixing and post-production workstation," in *Proc. 42nd Int. Computer Music Conference (ICMC)*, Utrecht, Netherlands, September 2016, pp. 122 – 127. http://hdl.handle.net/2027/spo.bbp2372.2016.023

[79] Holophonix by Amadeus. https://holophonix.xyz

[80] K. Hiyama, S. Komiyama, and K. Hamasaki, "The Minimum Number of Loudspeakers and its Arrangement for Reproducing the Spatial Impression of Diffuse Sound Field," in *Proc. 113rd Convention of the Audio Eng. Soc. (AES)*, Los Angeles, CA, USA, October 2002. www.aes.org/e-lib/browse.cfm?elib=11272

[81] A. Walther and C. Faller, "Assessing diffuse sound field reproduction capabilities of multichannel playback systems," in *Proc. 130th Convention of the Audio Eng. Soc. (AES)*, London, UK, May 2011. www.aes.org/e-lib/browse.cfm?elib=15895

[82] S. Agrawal and J. Braasch, "Impression of Spatially Distributed Reverberation in Multichannel Audio Reproduction," in *Proc. 145th Convention of the Audio Eng. Soc. (AES)*, New York, NY, USA, October 2018. www.aes.org/e-lib/browse.cfm?elib=19802

[83] C. Kirsch, J. Poppitz, T. Wendt, S. van de Par, and S. D. Ewert, "Spatial Resolution of Late Reverberation in Virtual Acoustic Environments," *Trends in Hearing*, vol. 25, pp. 1 – 17, 2021. https://doi.org/10.1177/23312165211054924

[84] J.-M. Jot, L. Cerveau, and O. Warusfel, "Analysis and Synthesis of Room Reverberation Based on a Statistical Time-Frequency Model," in *Proc. 103rd Convention of the Audio Eng. Soc. (AES)*, New York, NY, USA, September 1997. www.aes.org/e-lib/browse.cfm?elib=7150

[85] T. Carpentier, M. Noisternig, and O. Warusfel, "Hybrid Reverberation Processor with Perceptual Control," in *Proc. 17th Int. Conf. Digital*

Audio Effects (DAFx-14), Erlangen, Germany, September 2014, pp. 93 – 100. https://hal.science/hal-01107075/document
[86] J.-M. Jot, "An Analysis/Synthesis Approach to Real-Time Artificial Reverberation," in *1992 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1992. https://doi.org/10.1109/ICASSP.1992.226080
[87] J.-D. Polack, "Playing Billiards in the Concert Hall: The Mathematical Foundations of Geometrical Room Acoustics," *Applied Acoustics*, vol. 38, no. 2-4, 1993. https://doi.org/10.1016/0003-682x(93)90054-a
[88] D. Romblom, C. Guastavino, and P. Depalle, "Perceptual thresholds for non-ideal diffuse field reverberation," *J. Acoustical Society of America*, vol. 140, no. 5, pp. 3908 – 3916, November 2016. https://doi.org/10.1121/1.4967523
[89] D. Romblom, "Diffuse Field Modeling - The Physical and Perceptual Properties of Spatialized Reverberation," Ph.D. dissertation, McGill University, Montreal, Canada, 2016.
[90] B. Alary, P. Massé, S. J. Schlecht, M. Noisternig, and V. Välimäki, "Perceptual analysis of directional late reverberation," *J. Acoustical Society of America*, vol. 149, no. 5, pp. 3189 – 3199, 2021. https://doi.org/10.1121/10.0004770
[91] M. Nolan, E. Fernandez Grande, J. Brunskog, and C.-H. Jeong, "A wavenumber approach to quantifying the isotropy of the sound field in reverberant spaces," *J. Acoustical Society of America*, vol. 143, no. 4, pp. 2514–2526, 2018. https://doi.org/10.1121/1.5032194
[92] M. Berzborn and M. Vorländer, "Investigations on the Directional Energy Decay Curves in Reverberation Rooms," in *Proc. of Euronoise*, Crete, Greece, May 2018, pp. 2005 – 2010. www.euronoise2018.eu/docs/papers/337_Euronoise2018.pdf
[93] B. Alary, P. Massé, V. Välimäki, and M. Noisternig, "Assessing the anisotropic features of spatial impulse responses," in *Proc. EAA Spatial Audio Signal Processing Symposium*, Paris, France, September 2019, pp. 43 – 48. https://doi.org/10.25836/sasp.2019.32
[94] N. Meyer-Kahlen, S. J. Schlecht, and T. Lokki, "Parametric Late Reverberation from Broadband Directional Estimates," in *Proc. Int. Conf. Immersive and 3D Audio (I3DA)*, Bologna, Italy, September 2021. https://doi.org/10.1109/i3da48870.2021.9610928
[95] B. N. Gover, J. G. Ryan, and M. R. Stinson, "Measurements of directional properties of reverberant sound fields in rooms using a spherical microphone array," *J. Acoustical Society of America*, vol. 116, no. 4, pp. 2138–2148, 2004. https://doi.org/10.1121/1.1787525
[96] G. Götz, C. Hold, T. McKenzie, S. Schlecht, and V. Pulkki, "Analysis of multi-exponential and anisotropic sound energy decay," in *Proc. DAGA*, Stuttgart, Germany, March 2022.
[97] P. Massé, T. Carpentier, O. Warusfel, and M. Noisternig, "Measurement, analysis, and denoising of directional room impulse responses in complex spaces," in *Proc. Forum Acusticum*, Lyon, France, November 2020. https://hal.science/hal-03138302
[98] S. Oksanen, J. Parker, A. Politis, and V. Valimaki, "A directional diffuse reverberation model for excavated tunnels in rock," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 644 – 648. https://doi.org/10.1109/icassp.2013.6637727
[99] P. Massé, T. Carpentier, O. Warusfel, and M. Noisternig, "Denoising Directional Room Impulse Responses with Spatially Anisotropic Late Reverberation Tails," *Applied Sciences*, vol. 10, no. 3, pp. 1 – 17, February 2020. https://doi.org/10.3390/app10031033
[100] B. Alary and A. Politis, "Frequency-Dependent Directional Feedback Delay Network," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 176 – 180. https://doi.org/10.1109/icassp40776.2020.9054717
[101] C. Hold, T. McKenzie, G. Götz, S. J. Schlecht, and V. Pulkki, "Resynthesis of Spatial Room Impulse Response Tails With Anisotropic Multi-Slope Decays," *J. Audio Eng. Soc.*, vol. 70, no. 6, pp. 526 – 538, June 2022. https://doi.org/10.17743/jaes.2022.0017
[102] C. Kirsch, J. Poppitz, T. Wendt, S. van de Par, and S. D. Ewert, "Computationally Efficient Spatial Rendering of Late Reverberation in Virtual Acoustic Environments," in *Proc. Int. Conf. Immersive and 3D Audio (I3DA)*, Bologna, Italy, September 2021. https://doi.org/10.1109/i3da48870.2021.9610896
[103] R. Izhaki, *Mixing Audio - Concepts, Practices, and Tools*. Routledge, 2017. https://doi.org/10.4324/9781315716947
[104] M. A. Gerzon, "The Design of Distance Panpots," in *Proc. 92nd Convention of the Audio Eng. Soc. (AES)*, Vienna, Austria, March 1992. http://www.aes.org/e-lib/browse.cfm?elib=6825

APPENDIX
Spat_OPer PERCEPTUAL CONTROL EQUATIONS

A. Expression of the perceptual attributes

In this appendix, we review the set of equations resulting from the psycho-experimental study reported in section III-B. These equations express the acoustical criteria exposed in the high-level Spat_OPer control panel (Fig. 5 and Table I) as functions of the time-frequency-energy distribution in the impulse response (Fig. 6).

The impulse response energy is decomposed into four time sections as defined in Table III. The energy Ri of the i^{th} section (with $i \in \{0, 1, 2, 3\}$) is given by:

$$\mathrm{Ri} = \int_{t=\mathrm{li}}^{t=\mathrm{l(i+1)}} h^2(t) \cdot dt \,. \tag{1}$$

TABLE III: Temporal segmentation of the impulse response.

Section	Ri	li	l(i + 1)
Direct sound	R0	0 ms	20 ms
Early reflections	R1	20 ms	40 ms
Late reflections	R2	40 ms	100 ms
Reverberation tail	R3	100 ms	∞

We denote $\mathrm{Ri}_{\mathrm{low}}$, $\mathrm{Ri}_{\mathrm{high}}$, and $\mathrm{Ri}_{\mathrm{mid}}$ the energies in the low, high and mid frequency bands, respectively. Similarly, the variation of the reverberation decay time with frequency is represented by the values $\mathrm{Rt}_{\mathrm{low}}$, $\mathrm{Rt}_{\mathrm{high}}$ and $\mathrm{Rt}_{\mathrm{mid}}$ (also denoted Rt). This defines a total of 15 low-level rendering parameters. The psycho-experimental study reported in section III-C and [53] identified 10 mutually independent acoustical criteria which uniquely characterize the acoustical sensation perceived by a listener, produced by a sound source, both located at given positions in a concert hall or auditorium:

- The perception of the sound source is characterized by three criteria: the early energy Es and its relative variations at low and high frequencies denoted Desl and Desh.

Assuming that $\mathrm{R0}_{\mathrm{mid}} + \mathrm{R1}_{\mathrm{mid}} > \mathrm{R2}_{\mathrm{mid}}$,

$$\mathrm{Es} = \mathrm{R0}_{\mathrm{mid}} + \mathrm{R1}_{\mathrm{mid}} \,; \tag{2}$$

$$\mathrm{Desl} = \frac{\mathrm{R0}_{\mathrm{low}} + \mathrm{R1}_{\mathrm{low}}}{\mathrm{Es}} \,; \tag{3}$$

$$\mathrm{Desh} = \frac{\mathrm{R0}_{\mathrm{high}} + \mathrm{R1}_{\mathrm{high}}}{\mathrm{Es}} \,. \tag{4}$$

- The perception of the room is characterized by three criteria: the reverberation time Rt and its relative variations at low and high frequencies denoted Drtl and Drth.

$$\mathrm{Drtl} = \mathrm{Rt}_{\mathrm{low}}/\mathrm{Rt} \,. \tag{5}$$

$$\mathrm{Drth} = \mathrm{Rt}_{\mathrm{high}}/\mathrm{Rt} \,. \tag{6}$$

- The perception of source/room interaction is characterized by four acoustical criteria: Rd1, Rd2, Edt, Rev.

$$\mathrm{Rd1} = \frac{0.3 \cdot \mathrm{R1_{mid}} + 0.05 \cdot \mathrm{R2_{mid}}}{\mathrm{Es}} \,. \tag{7}$$

Assuming that $\mathrm{R0_{mid}} > \mathrm{R1_{mid}}$,

$$\mathrm{Rd2} = \frac{0.5 \cdot \mathrm{R1_{mid}} + 1.5 \cdot \mathrm{R2_{mid}}}{\mathrm{R0_{mid}}} \,. \tag{8}$$

Noting $C = 10^{\frac{-1.2}{\mathrm{Rt}}}$, $D = \mathrm{Es}/\mathrm{R3_{mid}}$ and $D_{\max} = 10^{\frac{3}{2}} - 1$.

If $\mathrm{Es} + \mathrm{R2_{mid}} \leq D_{\max} \cdot \mathrm{R3_{mid}}$, then:

$$\mathrm{Edt} = 0.4 + \mathrm{Rt} \cdot \left(1 - \frac{2}{3} \cdot \log_{10}\!\left(1 + \frac{\mathrm{Es} + \mathrm{R2_{mid}}}{\mathrm{R3_{mid}}}\right)\right) ; \tag{9}$$

otherwise:

$$\mathrm{Edt} = \frac{0.6}{\log_{10}\!\left(1 + \frac{\mathrm{Es}+\mathrm{R2_{mid}}}{\mathrm{R3_{mid}}}\right)} \,. \tag{10}$$

Lastly, if $(1 - C) \cdot \mathrm{R3_{mid}} \leq 4 \cdot \mathrm{Es}$ then:

$$\mathrm{Rev} = C \cdot \mathrm{R3_{mid}} + \frac{\left((1 - C) \cdot \mathrm{R3_{mid}}\right)^2}{8 \cdot \mathrm{Es}} \,; \tag{11}$$

otherwise: $\mathrm{Rev} = \mathrm{R3_{mid}} - 2 \cdot \mathrm{Es}$.

B. Derivation of the low-level rendering parameters

In this section, we review a set of equations previously reported in [15], [63], solving the inversion of the above relations in order to compute reverberator parameters given the settings of the acoustical criteria in the Spat_OPer control panel. For an omnidirectional sound source having flat axis and omni spectral corrections, the twelve energy values $\mathrm{Ri_{mid}}$, $\mathrm{Ri_{low}}$ and $\mathrm{Ri_{high}}$ (with $i \in \{0, 1, 2, 3\}$) are obtained by:

$$\begin{cases} \mathrm{R0_{low}} & = \mathrm{Desl} \cdot \mathrm{R0_{mid}} \,. \\ \mathrm{R0_{mid}} & = \mathrm{Es} - \mathrm{R1_{mid}} \,. \\ \mathrm{R0_{high}} & = \mathrm{Desh} \cdot \mathrm{R0_{mid}} \,. \end{cases} \tag{12}$$

$$\begin{cases} \mathrm{R1_{low}} & = \mathrm{Desl} \cdot \mathrm{R1_{mid}} \,. \\ \mathrm{R1_{mid}} & = (\mathrm{Es} \cdot \mathrm{Rd1} - 0.05 \cdot \mathrm{R2_{mid}})/0.3 \,. \\ \mathrm{R1_{high}} & = \mathrm{Desh} \cdot \mathrm{R1_{mid}} \,. \end{cases} \tag{13}$$

$$\mathrm{R2_{low}} = \mathrm{R2_{high}} = \mathrm{R2_{mid}} \,. \tag{14}$$

$$\mathrm{R3_{low}} = \mathrm{R3_{high}} = \mathrm{R3_{mid}} \,. \tag{15}$$

If $\mathrm{Edt} > 0.4$, then:

$$\mathrm{R2_{mid}} = \mathrm{R3_{mid}} \cdot \left(10^{\frac{3}{2} \cdot (1 + \frac{0.4 - \mathrm{Edt}}{\mathrm{Rt}})} - 1\right) - \mathrm{Es} \,; \tag{16}$$

otherwise:

$$\mathrm{R2_{mid}} = \mathrm{R3_{mid}} \cdot \left(10^{\frac{0.6}{\mathrm{Edt}}} - 1\right) - \mathrm{Es} \,. \tag{17}$$

If $\frac{\mathrm{Rev}}{\mathrm{Es}} \leq 2 \cdot \frac{(1+C)}{(1-C)}$ then:

$$\mathrm{R3_{mid}} = \left(-C + \sqrt{C^2 + \frac{\mathrm{Rev}}{2 \cdot \mathrm{Es}} \cdot (1 - C)^2}\right) \cdot \frac{4 \cdot \mathrm{Es}}{(1 - C)^2} \,; \tag{18}$$

otherwise: $\mathrm{R3_{mid}} = \mathrm{Rev} + 2 \cdot \mathrm{Es}$.

The acoustical criteria Rd1 and Rd2 are related respectively to the percepts of *apparent source width* (ASW) and *listener envelopment* (LEV) mentioned in section III-A. In order to control these two effects independently, it would be necessary to afford an additional degree of freedom in the low-level reverberation rendering model, enabling finer control of the spatial distribution of reflections. Instead, it was decided to discard Rd2 from the high-level control interface. The percept of listener envelopment (LEV) is therefore governed indirectly by adjustments of any of the criteria Es, Rd1, Edt, Rev or Rt, per Eq. (8) and Eq. (12)–(18).

In the implementation of the Room module, frequency-dependent gains are realized by the spectral correctors denoted Ri in Fig. 9, realized by proportional dual-shelving filters [68]. For a directional sound source, the omni spectral correction is applied by offsetting accordingly the low/mid/high dB gains in the spectral correctors R1, R2 and R3. The direct sound component R0 is corrected similarly, according to the sound source's axis and omni spectral correction settings, its orientation and its directivity pattern.

In the realization of this high-level control scheme, it is necessary to implement constraints ensuring that the values of R0, R1, and R2 remain always positive. Additionally, the maximum value of Rd1 is limited in order to prevent R0 from vanishing (since the direct sound arrival time provides the temporal reference on which the definition of all the acoustical criteria relies). Consequently, the allowed setting ranges of Rd1 and Edt depend on the settings of the other acoustical criteria, as indicated below.

$$\mathrm{Rd1_{min}} = 0.05 \cdot \mathrm{R2_{mid}}/\mathrm{Es} \,. \tag{19}$$

$$\mathrm{Rd1_{max}} = 0.27 + \mathrm{Rd1_{min}} \,. \tag{20}$$

If $2D \leq D_{\max}$, then:

$$\mathrm{Edt_{min}} = 0.4 + \mathrm{Rt} \cdot \left(1 - \frac{2}{3} \cdot \log_{10}(1 + 2D)\right) ; \tag{21}$$

otherwise:

$$\mathrm{Edt_{min}} = \frac{0.6}{\log_{10}(1 + 2D)} \,. \tag{22}$$

If $D \leq D_{\max}$, then:

$$\mathrm{Edt_{max}} = 0.4 + \mathrm{Rt} \cdot \left(1 - \frac{2}{3} \cdot \log_{10}(1 + D)\right) ; \tag{23}$$

otherwise:

$$\mathrm{Edt_{max}} = \frac{0.6}{\log_{10}(1 + D)} \,. \tag{24}$$