FIA 2020/22
XII CONGRESSO/CONGRESO IBEROAMERICANO DE ACÚSTICA
XXIX ENCONTRO DA SOCIEDADE BRASILEIRA DE ACÚSTICA - SOBRAC
Florianópolis, SC, Brasil

How do spatial audio plugins work and what functionalities do they offer: a comparative perspective

Rosero, K.[1]**; Abreu, D.**[1]**; Grijalva, F.**[2,3]**; Masiero, B.**[1]

[1] Faculdade de Engenharia Elétrica e de Computação (FEEC), Universidade Estadual de Campinas (UNICAMP), Campinas, SP, Brasil, k264373@g.unicamp.br, douglashsabreu@gmail.com, masiero@unicamp.br

[2] Faculty of Engineering and Applied Sciences (FICA), Telecommunications Engineering, Universidad de Las Américas (UDLA), Quito 170125, Ecuador, felipe.grijalva@udla.edu.ec

[3] Departamento de Electrónica, Telecomunicaciones y Redes de Información (DETRI), Escuela Politécnica Nacional, Quito 170109, Ecuador, felipe.grijalva@epn.edu.ec.

Abstract

Audio systems that reproduce spatial information are increasingly spreading across the market, aiming at offering high fidelity in the perception of timbres and high quality in the perception of spatial content. Special equipment, like microphones and loudspeaker arrays are commonly required for recording and reproducing spatial sound. Similar to traditional audio formats, the manipulation of immersive sound employs digital audio workstations (DAWs) with software extensions installed to expand the DAW's functionality, named "spatial audio plugins". As a result, it is necessary to disseminate among researchers and practitioners the most relevant spatial audio plugins available. Based on the production chain of immersive sound, we present plugins for the encoding and decoding stages with a comparative perspective of its features. Common operations provided by the available plugins encode regular audio formats such as mono, stereo or multi-channel sound into the Ambisonics spatial audio format of different orders. Decoding plugins process Ambisonic signals into binaural outputs or loudspeaker array formats. This paper summarizes the available spatial audio plugins and the recent developments they provide, briefly describing the parametric and non-parametric methods in which they are based. We hope this manuscript contributes to a better navigation through this large range of options regarding spatial audio plugins.

Keywords: Spatial audio, ambisonics, digital audio workstations, TV 3.0, MPEG-H.

PACS: 43.60.-c, 43.60.Ac, 43.60.Ek, 43.60.Dh.

1. INTRODUCTION

It is very common to use the terms Spatial Audio, Immersive Audio, 3D Audio, among others to describe the quality of an audio system that is able to recreate the spaciousness of a sound scene. In other words, any system that provides the listener with a sense of space beyond conventional stereo, allowing the listeners to identify where the sound is coming from, be it from above, below or 360 degrees around them, can be termed a spatial audio system. And, although the concept of spatial audio may seem to be something new, binaural[1] audio was invented already in the late 19th century [1].

Besides binaural technology, there is currently a large number of different methods, technologies and systems used to record, synthesize, encode, decode and reproduce spatial audio. And thanks to the rapid developments of audio for virtual reality applications, the implementation of spatial audio technology on major platforms such as Facebook and Youtube, and also the advancement of audio technologies — such as Dolby AC-4 [2] and MPEG-H [3] — aimed at the new generations of streaming and broadcasting television systems (TV 3.0) [4], the number of people interested in tools for manipulating these technologies has increased. In this context, one of the most important and frequently asked questions in the field of spatial audio is: "Where do I start?"

This article aims at describing, comparing and presenting available tools for manipulating spatial audio. The vast majority of the tools described are free, from virtual studio technologies (VST) to digital audio workstations (DAWs).

2. FUNDAMENTALS

To understand the tools described in this article, it is important to briefly review spatial hearing and also approaches related to spatial audio encoding and decoding.

[1]Term used to describe that two ears are involved in human hearing, the binaural concept is directly correlated with spatial hearing, and, therfore, spatial audio.

2.1 Spatial hearing

Sound waves arriving at the human ears are encoded by the auditory system to estimate, among other features, the nature and localization of their sources. In order to define the localization of a sound source, humans commonly use an egocentric reference system, in which the locations are specified by the azimuth and elevation angles, and the distance from the center of the head to the sound source [5].

As the human ears are positioned on opposite sides of the head, the differences in the sound pattern perceived by each ear help us to detect and localize sounds. The difference in the time of arrival to each ear is known as interaural time difference (ITD), and the different sound intensity that each ear perceives is known as the interaural level difference (ILD), and both are called binaural cues as they depend on both ears [6]. Moreover, the individual structure of the ear pinna results in a direction-dependent spectral shaping of the incoming sound. Both monaural and binaural cues are the information that the auditory system encodes from acoustical environments, which are naturally three-dimensional. The aforementioned cues can be modeled by a set of functions called Head-Relative Transfer Functions (HRTFs) [7], which describe the spectral filtering over a sound source caused by the head, pinna and torso before it reaches the eardrum.

The most common real-life sound environments in which we live are rooms, which reflect the sound in the walls, ceiling and floor. The behavior of sound in rooms can be characterized by the room impulse responses (RIRs). Therefore, the synthesis of three-dimensional virtual sound spaces is one of the basic application to offer an immersive sound experience. In contrast, spatial audio recording, processing and playback techniques focus on maintaining the spatial audio properties, which can be perceived by humans as in natural hearing. In the next section, we will explain the workflow of spatial audio production.

2.2 Spatial audio

Spatial audio technologies have gained attention principally due to improvements in digital signal processing techniques and audio hardware development, designed to enhance the spatial quality of reproduced sound [8]. Since the first stereo transmission in 1881, sound industry has been interested in offering some spatial realism in sound reproduction.For instance, the cinema stereo evolution developed surround sound audio systems as 5.1, 7.1, 10.2, 11.1, 22.2, which define loudspeakers positions to create an immersive sound experience. Further information about surround systems can be found in [9, 10].

According to the ISO/MPEG-H standardization group [3], spatial sound acquisition and subsequent processing, transmission and reproduction follow three principal approaches: channel-based, object-based, and scene-based format. As they are not mutually exclusive, they can be used together. Next, these approaches will be briefly explained.

2.2.1 Channel-based approach

This approach renders spatial information through the difference between the transmitted waveforms, called *channel signals*. Each input audio channel feeds a loudspeaker placed in an intended position relative to the listener. In this way, the renderer or decoder retrieves the produced audio channels and feed them to the corresponding loudspeakers [11]. The placement and mixing of the channels depend on the application of digital perceptual cues, such as ITDs and ILDs.

2.2.2 Object-based approach

As explained in [11], this approach describes the spatial sound scene by a number of uncorrelated sound sources, also called *sound objects*, such that each sound can be placed in a specific position of the auditory scene, independent of the availability of loudspeaker position. Hence, sound objects generally need to be rendered to their target positions by algorithms as Vector Base Amplitude Panning (VBAP) [12] or Multiple Direction Amplitude Panning (MDAP) [13]. The processing module prepares the objects along with the scene description. Hence, several parameters are needed for successful rendering, such as position, orientation, and reverberation parameters.

In this context, the acoustic scene parameters can be obtained by using parametric models. Once the sound field is analyzed, the parameters of the sound scene can be transmitted, manipulated, and synthesized [14]. The parameters' extraction is achieved by analyzing the sound field in narrow frequency bands using several microphones. Then, instead of transmitting many microphone signals, only the direct and diffuse components are transmitted along with the parametric information. Examples of parametric models include: higher order directional audio coding (HO-DirAC) [15], COding and Multidirectional Parametrization of Ambisonic Sound Scenes (COMPASS) [16], etc. Further information about parametric models can be found in [14].

Parametric models are not exclusively used by the object-based approach, these are also useful techniques for the scene-based approach that will be explained in the following section.

2.2.3 Scene-based approach

This approach uses the Ambisonics format to describe the spatial sound scene. Based on principles of psychoacoustics, Ambisonics was first presented as a unified system for directional sound pickup, storage or transmission, and reproduction. The sound is encoded from all directions in terms of pressure and velocity components by using microphone arrays, and these signals are decoded to different loudspeaker configurations [9]. The first-order Ambisonics (FOA) consist of an omnidirectional signal W, and three directional signals (X, Y, Z) aligned with the Cartesian coordinate axes [17]. This is called the B-format of Ambisonics.

The incorporation of more recording channels can improve directional encoding, as it provides more coefficient signals for increasing spatial selectivity. These second and higher order Ambisonics (HOA) are expressed in terms of the real spherical harmonics (SH) [18],

$$Y_n^m(\phi,\theta) = N_n^{|m|} P_n^{|m|} \sin\theta \cdot \begin{cases} \sin(|m|\phi) \text{ if } m<0, \\ \cos(|m|\phi) \text{ if } m\geq 0, \end{cases} \tag{1}$$

where Y denotes the spherical harmonic of order n and degree m with ranges of $-n \leq m \leq n$, and $0 \leq n \leq N$. N_n^m is a normalization factor, P_n^m is the associated Legendre polynomial, ϕ the azimuth angle, and θ the elevation angle. Further information about SHs can be found in [19].

One drawback of the Ambisionics format is the existence of multiple standards concerning normalization and chaneel ordering. The traditional B-format, which consists of four channels, was expanded by the Furse-Malham format (FuMa), such that channels up to the third Ambisonics order could be identified. For HOA, the Ambisonic Channel Number (ACN) format was considered, as it orders the channels corresponding to the SHs following the relation $ACN = l^2 + l + m$. Regarding the Ambisonics' normalization techniques, the most common are: 1) maxN, which is used in the FuMa format and is compatible with the B-format, 2) SN3D semi-normalization, adopted for the ambiX format [18], and 3) N3D normalization, which is also supported by several Ambisonics applications and has a simple relation with SN3D. Authors in [18] proposed ambiX, a standard Ambisonics format that facilitates exchange and communication when using Ambisonics. This considers ACN as the official channel ordering system, and SN3D the normalization factor. The ambiX format has been widely implemented in spatial audio plugins.

3. SPATIAL AUDIO PLUGINS

Creating and manipulating spatial audio requires digital processing of the audio signals, which can be managed by audio plugins. Software developers and audio researchers have provided computational tools that help us synthesize, or process, spatial sound. Virtual Studio Technology (VST) and Avid Audio Extension (AAX) audio plugins designed to handle spatial audio can be incorporated in DAWs that allow multichannel inputs.

Spatial audio recordings commonly rely on an array of microphones that capture the entire surrounding sound scene. Nowadays, several commercial microphone arrays are in the market, such as: Eigenmike, Zylia, VisiSonics, Zoom H3-VR, etc. The recorded audio signals are commonly processed using the scene-based approach, resulting in a greater number of plugins based on Ambisonics. However, the parameters' extraction of the sound scene is also achieved by using parametric models.

The maximum Ambisonics order N that the plugins accept, is generally limited by the maximum number of channels that the host DAW can support. However, plugins that rely on parametric models are typically limited to the third Ambisonics order due to its high computational cost. Table 1 presents the plugin suites explored in this work, as well as the maximum order that they can deal with, and whether they are free or licensed.

Table 1: Spatial audio plugin suites.

Plugin suite	Order	Availability
IEM	7th	Free
ambiX	7th	Free
mcfx	7th	Free
SPARTA	7th	Free
COMPASS	3rd	Free
HO-DirAC	3rd	Free
Blue Ripple Sound	3rd	Licensed
b<>com	4th	Licensed
HARPEX	3rd	Licensed
SSA	7th	Licensed
dearVR Ambi Micro	3rd	Free
dearVR Pro	3rd	Licensed

In order to analyze and compare the functionalities offered by the studied spatial audio plugins, we separate them into these categories: encoding plugins, decoding plugins, rendering plugins, and digital spatial audio effects.

3.1 Encoding plugins

The encoding plugins aim to transform mono or multichannel audio signals into different orders of Ambisonics. The explored plugins are listed in Table 2, in which we categorized them to compare their characteristics.

For panning the audio sources, it is necessary to define the desired sound position in terms of azimuth and elevation angles, from which the listener will perceive the sound coming from. `IEM StereoEncoder` and `SSA aXPanner` plugins are the most basic tools, which are able to generate spatial sound up to 7th order of

Table 2: Spatial audio plugins for encoding purposes.

Plugin \ Feature	Input	Output	Panning	Reverberation	Filter	Directivity
IEM - StereoEncoder	Mono, Stereo	ambiX - 7th	X			X
IEM - DirectivityShaper			X		X	X
SSA - aXPanner			X			
IEM - Multiencoder	64 ch		X			
ambiX - Encoder			X			X
SPARTA - AmbiENC			X			
b<>com - HOAPan	8 ch	ambiX - 3rd	X			
O3A - Panner			X			
IEM - RoomEncoder	Mono	ambiX - 7th	X	X		
SPARTA - 6DoFconv			X	X		
SPARTA - AmbiRoomSim	64 ch		X	X		
COMPASS - 6DoF*	16 ch	ambiX - 3rd	X	X		
SPARTA - Array2SH	Microphone array signals	ambiX - 7th			X	
ATK - AtoB		ambiX - 1st				

*Parametric processing

Ambisonics by panning static sound sources. IEM DirectivityShaper offers the same functionality, however, the input sound can be filtered into four independent frequency bands, each one with its own filter parameters and Ambisonics order. Moreover, the radiation pattern (or directivity) of the sound source can be adjusted. Among the plugins able to encode and panning more than two sounds we found: IEM MultiEncoder in which the gain of each source can be also modified, ambiX Encoder that also offers directivity adjustment, and SPARTA AmbiENC that includes several loudspeaker layout positions as presets, in order to facilitate the panning alignment of the sound sources. The b<>com HOAPan and O3A Panner panning function supports up to 8 independent input sound sources, which can be encoded into the 3rd Ambisonics order. In b<>com HOAPan, a distance attenuation and a near-field effect can also be added to the sounds. These encoding plugins consider a far-field condition.

In addition to panning a sound source, IEM RoomEncoder is a room acoustic simulator, in which the user can move a sound source and a listener into a virtual shoebox-shaped room, simulating over 200 wall reflections. Attenuation parameters can be also adjusted as well as the room size. The SPARTA AmbiRoomSim plugin also simulates room reflections, but it allows up to 64 sound sources that can be placed in the virtual room. This plugin's output can be a single 7th order Ambisonics signal, or several lower order signals, as long as the 64 channels are not exceeded. In contrast, external spatial room impulse responses (SRIRs) datasets can be imported in SPARTA 6DoFconv plugin to simulate how a sound source would be perceived by the listener in different positions of the room and with different head orientations.

Based on [20], SPARTA Array2SH plugin encodes spherical/cylindrical array signals into Ambisonics format by using analytical descriptors, which determine the influence of the microphone array physical properties on the incoming sound. Depending on the microphone array characteristics, the plugin estimate the order-dependent equalization curves that remove the influence of the array itself. Besides including the specifications for several commercially available microphone arrays, it is a convenient mean for obtaining the spherical harmonic signals for custom-build arrays. Lastly, the ATK AtoB plugin encodes four microphone signals into the 1st order of Ambisonics by equally distributing the sound throughout the sound field.

The plugins explained in this section output Ambisonic signals, which need to be decoded for playback. In the next section, we present several decoding plugins available for using within DAWs.

3.2 Decoding plugins

One goal of the research in the audio reproduction field is bringing a three-dimensional sound field to a listening situation [12], either virtually created acoustic scenarios or recorded scenarios. This section includes several plugin descriptions for decoding Ambisonic signals into mono, stereo, binaural output, or into a loudspeakers array. Table 3 compares the principal features of the analyzed plugins.

3.2.1 Loudspeaker layout decoders

For decoding Ambisonic signals into specific loudspeaker arrangements, `ambiX Decoder` and `b<>com HOA2Spk` rely on a set of configuration files containing a decoder matrix. In that way, several standard surround configurations can be selected to playback spatial sound. The `IEM SimpleDecoder` plugin also allows the user to import personalized configuration files to decode the Ambisonic input. It also offers a subwoofer functionality that allows us to separate the input signal into two frequency bands. When using the discrete subwoofer, the upper band follows the configuration decoding file, and the lower band is sent to a defined subwoofer channel; whereas, the virtual subwoofer passes the lower band to all loudspeakers. The `IEM AllRADecoder` is a powerful plugin which supports all the creation process of an Ambisonic decoder for any desired loudspeaker layout. Once the user has defined the number of loudspeakers and its positions, the plugin uses the AllRAD technique described in [21] to triangulate the layout and even include imaginary loudspeakers to disambiguate certain layouts. The Ambisonics order is suggested to be chosen depending on the separation between the loudspeakers, for example: 2nd order for 60° spacing, or 7th order for 15°. In the final stage, the decoder can be directly applied to Ambisonics input signals, or exported to a JSON file, which in turn can be used in plugins that accept an imported configuration file, as `IEM SimpleDecoder`.

3.2.2 Binaural decoders

`IEM BinauralDecoder` uses the Magnitude Least-Squares (MagLS) technique to decode the Ambisonic input into binaural headphone signals. The MagLS technique designs binaural rendering filters for Ambisonic signals based on magnitude-only optimization at high frequen-

Table 3: Spatial audio plugins for decoding purposes.

Plugin \ Feature	Input	Output	HRTFs	HT	Filter	Diffuseness	DOA
IEM - AllRADecoder	ambiX - 7th	Loudspeakers layout					
IEM - SimpleDecoder					X		
ambiX - Decoder							
b<>com - HOA2Spk	ambiX - 4th						
IEM - BinauralDecoder	ambiX - 7th	Binaural	X				
ambiX - Binaural	ambiX - 7th		X				
SPARTA - AmbiBIN	ambiX - 7th		X	X		X	
SSA - aXMonitor	ambiX - 7th		X				
b<>com - HOA2Bin	ambiX - 4th		X				
COMPASS - Binaural*	ambiX - 3rd		X	X		X	X
COMPASS - Binaural VR*	ambiX - 3rd		X	X			X
CroPaC - Binaural*	ambiX - 1st		X	X		X	X
SPARTA - AmbiDEC	ambiX - 7th	Binaural, Loudspeakers layout	X				
COMPASS - Decoder *	ambiX - 3rd		X			X	X
O3A - Decoder	ambiX - 3rd		X				
ATK - Decoders	ambiX - 1st	Mono, stereo, binaural, loudspeakers layout	X				
O3ABeamer - Virtual Microphone	ambiX - 3rd	Mono, stereo					
IEM - ProbeDecoder	ambiX - 7th	Mono					

*Parametric processing

cies, and disregarding the interaural phase differences above 2 kHz [22]. Moreover, it includes headphone equalization for several commercial headphones. In the `ambiX Binaural` plugin, the Ambisonic signals are convolved with the associated binaural loudspeaker IRs, this plugin employs virtual loudspeakers. Using the `SPARTA AmbiBIN` plugin, the user can select from the following decoding methods: Least-squares (LS), spatial re-sampling (SPR) [23], MagLS, or time-aligment [24]; the last mentioned method can apply diffuseness correction. This plugin also accepts, as input, positional information from a head tracker via an OSC connection. A simpler binaural decoding plugin that still maintains the maximum 7th order of Ambisonics is `SSA aXMonitor`, which is commonly used to monitor how the Ambisonic signals sound over headphones. This plugin convolves the input signals with a set of HRTFs, that can also be loaded by the user via SOFA files. The `b<>com HOA2Bin` plugin includes, as presets, different HRTFs sets; some of them even belong to a reverberant acoustic scene.

We found several binaural decoders based on parametric techniques. They usually manage up to 3rd order Ambisonics, and can involve more computational complexity. The COMPASS method, used in the `COMPASS Binaural` plugin, operates on time-frequency transformed signals, and processes the sound field to estimate its direct and diffuse sound components. It can extract multiple source components up to half the number of input channels by using MUSIC [25] or ESPRIT [26] DoA estimators. The direct sound component DoAs are used to process the binaural output [16]. Then, the source components are transmitted to the channels with maximum directional concentration from their estimate DoA, following the amplitude panning gains, or the HRTFs. The user can import personalized HRTFs via SOFA files, and communicate with a head tracking device via OSC messages. Moreover, the `COMPASS Binaural VR` version of this plugin also supports listener translation.

Binaural decoding of FOA signals can be achieved by the `CroPaC` plugin [27], which offers a performance comparable with 3rd order Ambisonics because of its high perceptual accuracy. `CroPaC` estimates the sound source DoA via steered-response power (SRP) beamforming [28] and segregates the components founded in that localization. Then, the ambient stream and the sound source components are binauralized according to the provided (or imported) HRTFs. It also supports head tracking via OSC messages.

3.2.3 Decoding plugins with several output options

The `COMPASS Binaural` method was included in the `COMPASS Decoder` plugin, that processes Ambisonic signals for loudspeakers playback and binaural reproduction. In the same fashion, the `SPARTA AmbiDEC` plugin contains `SPARTA AmbiBIN`, offering a binaural output as well as loudspeaker layouts playback. The principal characteristic that makes this plugin a more powerful tool is the inclusion of three decoding techniques: AllRAD decoder, Energy-Preserving Ambisonic Decoder (EPAD) [29], and Mode-Matching Decoding (MMD) [30], which can be applied for low or high frequencies depending on the necessities. Furthermore, each channel can be panned in a specific position.

As O3A plugins are licensed, `O3A Decoder` uses patented techniques to process the Ambisonic signals into a binaural output or loudspeaker arrays. Additional details about the decoding process are no available. The `O3A Beamer Virtual Microphone` extracts sound in a particular direction defined by azimuth and elevation angles, and the user can control the sharpness factor. The output signal can be mono or stereo, hence, this plugin can be useful to monitor a specific sound source in Ambisonic signals. Likewise, the `IEM ProbeDecoder` offers a directional mono channel output useful for testing purpose. Finally, `ATK Decoders` are restricted to process just FOA signals, and by using virtual microphone decoders, extracts mono, stereo, binaural, and loudspeaker array signals.

3.2.4 Ambisonics upmixing

The order of the Ambisonic signals directly influences the spatial resolution of the sound scene. Physical and practical limitations in the recording microphone array restrict the maximum order at which the sound scene may be acquired [31], which consequently hampers the accuracy of the spatial audio playback. Aiming to alleviate this difficulty, several plugins upmix lower-order Ambisonic signal into HOA. Offering a basic upmixing functionality, `O3A Injector` and `HARPEX ambiX` process up to 2nd order Ambisonic signals and decode it into 3rd order. `COMPASS Upmixer` and `HO-DirAC Upmixer` were implemented using the COMPASS approach to extract the Ambisonic recording parameters, and upmix the signals up to 7th order. The user can also modify the diffuseness parameter, which controls the diffuse-to-direct balance per frequency band. The more ‘direct’ sound is selected, the more the ambient component would be attenuated, and the sound source emphasized.

Some other plugins include the upmixing tool, but also offer additional functionalities. In the `COMPASS Gravitator` plugin, the user defines Cartesian coordinates from which the DoA is estimated. Later, the ’gravity’ parameter controls if the enhanced Ambisonics order is applied equally on the sound source and the auditory scene parameters, or just on the sound source [32]. On the other hand, the principal objective of the `COMPASS SideChain` plugin is to extract the sound scene parameters from one Ambisonic sound scene, and apply those parameters in the synthesis of a second Ambisonic scene, with the option of exporting a higher order signal up to 7th order. Lastly, in the `COMPASS SpatEdit` plugin, the user first define markers that can follow the sound sources’ direction. The signals extracted by the beamformer can be modified with audio effects or gain modification, and then they are passed to the second instance of the plugin, in which the signals are upmixed into HOA. Further details can be found in [33].

3.3 Rendering plugins

Some spatial audio plugins include encoding and decoding stages. In this work, we will refer to them as ‘renderers’ because they receive audio signals as input, and output playback audio formats instead of Ambisonic signals. In this section, we describe and compare the rendering plugins functionalities.

`SPARTA Binauraliser`, `b<>Spk2Bin`, and `HO-DirAC Binaural` receive up to 64, 32, and 16 audio channels respectively, and render them into binaural signals. Regarding specific features of each plugin, `SPARTA Binauraliser` convolves input audio with the HRTFs by applying normalized VBAP gains to the HRTF magnitudes and ITDs. Additionally, the sound sources can be individually panned and filtered. `b<>Spk2Bin` uses a virtual loudspeaker approach as well as HRTFs with or without reverberation effects, and `HO-DirAC Binaural` bypasses loudspeaker rendering by using HRTFs directly; it also can receive head tracking messages by an OSC port.

Similar to `SPARTA Binauraliser`, `SPARTA Panner` applies the VBAP method to the frequency-dependent panned input signals. However, this renders the spatialized audio into loudspeaker layout presets. `SPARTA Spreader` is also a renderer, which includes the additional function of spreading an input sound source to be reproduced over a large spreading area in a diffuse manner. This is achieved by creating mutually incoherent copies of the input signals and applying decorrelators, as explained in [34].

`HADES` plugin processes 8 head-worn microphone array signals and renders them into a binaural output. An initial estimation of the binaural signals can be achieved by: 1) a channel-based option, which select two reference channels to be routed into the binaural channels, 2) filter-and-sum beamformers [35], or 3) Binaural Minimum-Variance Distortionless Response (BMVDR) beamformers [36]. Then, the plugin applies the spatial covariance matching framework [37] to enhance the initial estimate of binaural signals. Also, with respect to head-worn microphone arrays, the `UltraSonic Super Hearing` plugin renders the 6 input signals into a binaural output, with the special feature of making ultrasonic sources audible for humans [38]. It employs the DoA estimation and the pitch is shifted to an audible range.

Among the rendering plugins which input is limited to 16 channels, we found:

1) `HARPEX Stereo`, in which the directivity pattern of the virtual microphones can be adjusted before being passed to the stereo output, 2) `HARPEX Surround`, in which the inputs can be panned to be reproduced by a horizontal loudspeaker array, 3) `dearVR Pro` that can include reverberation, and the input signals are panned into loudspeaker array presets, and 4) `HO-DirAC Decoder` that renders commercial microphone array signals by estimating loudspeaker signals using the EPAD technique, and comparing it with the parametrical analysis of the sound scene [39]. This can also produce a binaural output, however, due to the processing time, it can not be considered a lower-latency time plugin.

Lastly, four rendering plugins limit their input to 4 channels, and render them into mono, binaural, stereo, or basic loudspeaker layouts. `Soundfield by RØDE` plugin is exclusively related to the NT-SF1 Ambisonics microphone, and `dearVR Ambi Micro` to the Sennheiser AMBEO VR microphone. `HARPEX Shortgun` input can be raw tetrahedral or Double MS microphones to be rendered.

3.4 Digital spatial audio effects

Spatial audio can be manipulated to achieve a series of audio effects that modify the directionality, spaciousness, reverberance, etc., of the sound scene. In this section, we present several plugins that achieve spatial audio effects and can also be included in DAWs. The Ambisonics order supported by each digital effect corresponds to the maximum order of each plugin suite (Table 1).

Scene rotators: The Ambisonic scene can be rotated around the coordinate axes, across an arbitrary plane, or around the origin. The plugins associated to this feature are: `IEM SceneRotator`, `ambiX Rotator`, `ambiX widening`, `SPARTA Rotator`, `O3A Rotation`, `HARPEX Sound field rotation`, `SSA axRotate`, `ATK Mirror`, and `ATK RotateTiltTumble`. Related with this feature, `IEM Coordinate Converter` enables to switch between spherical or Cartesian coordinates, and `ambiX warp` that distorts the sound field towards a plane.

Transformation matrix: The user can load a configuration file containing a matrix that will be applied to the input signals. This could be used to apply spatial reverberation, for encoding microphone array signals for which an encoding matrix is available, even to obtain a binaural output by using a binaural head-related impulse response matrix. For these spatial audio transformations, we refer the reader to `IEM MatrixMultiplier`, `mcfx convolver`, and `SPARTA MatrixConv` plugins.

Filters/compressors: Some plugins offer multichannel equalizers with different number and type of filters applied to the frequency bands of the input signals. Among the spatial plugins with these characteristics, we found: `IEM MultiEQ`, `SPARTA MultiConv`, `mcfx filter`, `SPARTA Decorrelator`, `SSA aXDeesser` and `SSA aXEqualizer`. Several plugins modify the gain of the input signals by incrementing or compressing the amplitude, such as: `IEM OmniCompressor`, `IEM MultiBandCompressor`, `mcfx gain delay`, `O3A Gain`, `SSA aXCompressor`, `SSA aXGate`. With the objective to filter specific parts of the sound field, `IEM DirectionalCompressor`, `ATK Dominance` and `ambiX DirectionalLoudness` amplify, filter or attenuate a region of the input signals.

Distance effects: The `IEM DistanceCompensator` and `ATK Near-field` plugins calculate the needed delays and gains to compensate for distance differences between loudspeakers, or reduce the proximity effect of encoded signals.

3.5 Audio plugins for television systems 3.0

The Advanced Television Systems Committee (ATSC) is an international organization that develops standards for digital television. Hence, the ATSC is responsible for defining the next-generation broadcast standard for television. ATSC 3.0 audio standards (A/342 Part 1, "Audio Common Elements") offer to bring sound to a full 7.1+4 implementation, and incorporate both Dolby AC-4 (A/342 Part 2, "AC-4 System") and MPEG-H audio (A/342 Part 3, "MPEG-H

System") into its audio encoder and decoder recommendations [40].

3.5.1 Dolby Audio System (AC-4) features

The Dolby Encoding Engine is a file-based command line interface application for encoding Dolby Audio and Dolby Vision content [41]. The engine includes a robust CLI and an API that allows for straightforward integration into third-party products and workflows.

Figure 1 depicts a Dolby emission encoding system, which consists of an AC-4 encoder combined with input and output stages appropriate to the application. The emission encoding system is not a normatively-specified part of the AC-4 standard. The user can request an evaluation version of this licensed tool.

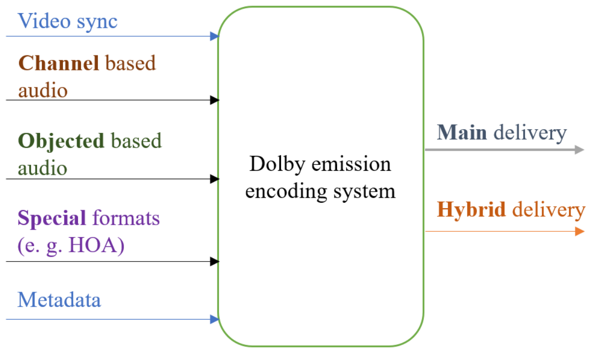

Figure 1: Dolby Emission Encoding System.

3.5.2 MPEG-H audio system features

MPEG-H audio is the only Next Generation Audio (NGA) codec able to provide immersive sound using any combination of the three well-established audio formats: channel-based, object-based, and scene-based. It is possible to create audio metadata to be used as input to an encoder through the MPEG-H Authoring Suite (MAS), which is an open-source and freely available software [42] created by Fraunhofer IIS for authoring MPEG-H audio scenes.

MAS is a set of tools, among which the MPEG-H Authoring Tool (MHAT) stands out. The tool enables easy MPEG-H authoring without the need for a DAW. You can define specific MPEG-H parameters, instantly listen to your configurations, and export your authored mixes as MPEG-H Production Format (MPF), MPEG-H BWF/ADM, or as a template XML file. Among its features we find: monitoring layouts up to 22.2, combination of channel-based components and dynamic audio objects, interactivity features for component gain and position, and binaural monitoring for immersive audio reproduction over headphones.

Another tool to be mentioned is the MPEG-H Conversion Tool (MCO), which converts MPEG-H content between different file formats. The MCO serves as an interface to the MPEG-H Audio ecosystem, it also allows importing and exporting MPEG-H Master files, as well as BWF/ADM personalized files created by other tools.The resulting files can even be edited and enhanced using the MAS advanced feature tools. In that way, MPEG-H authoring allows the conversion of object-only content for music productions, and can be integrated into most existing workflows that support up to 32 channels.

4. FINAL REMARKS

The development of technological devices that support spatial audio has gained the interest of both the public and the major technology companies. Hence, there is a need to explore spatial audio plugins that help users, small producers, and developers, to manage or even generate spatial audio. The free or licensed plugins presented in this work cover most of the current needs for spatial audio manipulation, from the encoding of microphone signals, through the application of audio effects, to the decoding into different playback systems.

We hope that a greater number of people and companies, who already work or intend to work with these technologies, find this article an enlightening way to apply the spatial audio plugins that are spread across the web. Additionally, the presented comparative approach aims to ease the selection of the most suitable audio plugins.

ACKNOWLEDGMENTS

This work was supported by the São Paulo Research Foundation (FAPESP), under grants #2017/08120-6 and #2019/22945-3.

REFERENCES

[1] Stephan Paul. Binaural recording technology: A historical review and possible future developments. *Acta acustica united with Acustica*, 95(5):767–788, 2009.

[2] Kristofer Kjörling, Jonas Rödén, Martin Wolters, Jeff Riedmiller, Arijit Biswas, Per Ekstrand, Alexander Gröschel, Per Hedelin, Toni Hirvonen, Holger Hörich, et al. Ac-4–the next generation audio codec. In *Audio Engineering Society Convention 140*. Audio Engineering Society, 2016.

[3] Jürgen Herre, Johannes Hilpert, Achim Kuntz, and Jan Plogsties. Mpeg-h 3d audio—the new standard for coding of immersive spatial audio. *IEEE Journal of Selected Topics in Signal Processing*, 9(5):770–779, 2015. doi:10.1109/JSTSP.2015.2411578.

[4] Rich Chernock, David Gomez-Barquero, Jerry Whitaker, Sung-Ik Park, and Yiyan Wu. Atsc 3.0 next generation digital tv standard—an overview and preview of the issue. *IEEE Transactions on Broadcasting*, 62(1):154–158, 2016.

[5] Stephen M Town, W Owen Brimijoin, and Jennifer K Bizley. Egocentric and allocentric representations in auditory cortex. *PLoS Biol*, 15(6):e2001878, June 2017. doi:10.1371/journal.pbio.2001878.

[6] William Yost and Raymond Dye. Fundamentals of Directional Hearing. *Seminars in Hearing*, 18: 321–344, 11 1997. doi:10.1055/s-0028-1083035.

[7] Henrik Møller, Michael Friis Sørensen, Dorte Hammershøi, and Clemen Boje Jensen. Head-Related Transfer Functions of Human Subjects. *Journal of the Audio Engineering Society*, 43(5):300–321, May 1995.

[8] Y. Suzuki, D. Brungart, Y. Iwaya, K. Iida, D. Cabrera, and H. Kato. *Principles and Applications of Spatial Hearing*. World Scientific Publishing Co. Pte. Ltd., Danvers, 2011. ISBN 981-4313-87-4.

[9] Francis Rumsey. *Spatial Audio*, volume 2nd of *Music Technology Series*. Focal Press, London, England, 1st edition, 2001. ISBN 0240516230.

[10] Silva, R. Surround Sound: The Audio Side of Home Theater. https://www.lifewire.com/surround-sound-home-theater-1846357, December 2021.

[11] Chungeun Kim. *Object-Based Spatial Audio: Concept, Advantages, and Challenges*, pages 79–84. Springer New York, New York, NY, 2014. ISBN 978-1-4614-8373-1. doi:10.1007/978-1-4614-8373-1_5.

[12] Ville Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6):456–466, June 1997.

[13] Ville Pulkki. Generic panning tools for max/msp. In *Proceedings of International Computer Music Conference*, page 304307, 2000.

[14] Konrad Kowalczyk, Oliver Thiergart, Maja Taseska, Giovanni Del Galdo, Ville Pulkki, and Emanuel AP Habets. Parametric spatial sound processing: A flexible and efficient solution to sound scene acquisition, modification, and reproduction. *IEEE Signal Processing Magazine*, 32(2):31–42, 2015.

[15] Ville Pulkki, Symeon Delikaris-Manias, and Archontis Politis. *Higher-Order Directional Audio Coding*, pages 141–159. 2018. doi:10.1002/9781119252634.ch6.

[16] Archontis Politis, Sakari Tervo, and Ville Pulkki. Compass: Coding and multidirectional parameterization of ambisonic sound scenes. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6802–6806. IEEE, 2018.

[17] Franz. Zotter. *Ambisonics A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality* . Springer Topics in Signal Processing, 19. Springer International Publishing, Cham, 1st ed. 2019. edition, 2019. ISBN 3-030-17207-4.

[18] C. NACHBAR, F. ZOTTER, E. DELEFLIE, and A. SONTACCHI. *ambix* - a suggested ambisonics format. In *Proceedings of the Ambisonics Symposium 2011*, Lexington KY, July 2011.

[19] Boaz Rafaely. *Fundamentals of Spherical Array Processing*. Springer Nature Switzerland AG, 2019. ISBN 1866-2609. doi:10.1007/978-3-319-99561-8.

[20] Leo McCormack, Symeon Delikaris-Manias, Angelo Farina, Daniel Pinardi, and Ville Pulkki. Real-time conversion of sensor array signals into spherical harmonic signals with applications to spatially localized sub-band sound-field analysis. In *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

[21] Franz Zotter and Matthias Frank. All-round ambisonic panning and decoding. *Journal of the Audio Engineering Society*, 60(10):807–820, October 2012.

[22] Christian Schörkhuber, Markus Zaunschirm, and Robert Höldrich. Binaural rendering of ambisonic signals via magnitude least squares. In *Fortschritter der Akustik (DAGA) 2018*, March 2018. ISBN 978-3-939296-13-3.

[23] Leo McCormack and Symeon Delikaris-Manias. Parametric first-order ambisonic decoding for headphones utilising the cross-pattern coherence algorithm. In *EAA Spatial Audio Signal Processing Symposium*, pages 173–178, Paris, France, September 2019. doi:10.25836/sasp.2019.26.

[24] Markus Zaunschirm, Christian Schörkhuber, and Robert Höldrich. Binaural rendering of ambisonic signals by head-related impulse response time alignment and a diffuseness constraint. *The Journal of the Acoustical Society of America*, 143(6):3616, June 2018. doi:10.1121/1.5040489.

[25] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986. doi:10.1109/TAP.1986.1143830.

[26] R. Roy and T. Kailath. Esprit-estimation of signal parameters via rotational invariance techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(7):984–995, 1989. doi:10.1109/29.32276.

[27] Symeon Delikaris-Manias and Ville Pulkki. Cross pattern coherence algorithm for spatial filtering applications utilizing microphone arrays. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2356–2367, 2013. doi:10.1109/TASL.2013.2277928.

[28] Jacek P. Dmochowski, Jacob Benesty, and SofiÈne Affes. A generalized steered response power method for computationally viable source localization. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2510–2526, 2007. doi:10.1109/TASL.2007.906694.

[29] Franz Zotter, Hannes Pomberger, and Markus Noisternig. Energy-preserving ambisonic decoding. *Acta Acustica united with Acustica*, 98:37–47, 01 2012. doi:10.3813/AAA.918490.

[30] Tianshu Qu, Huang Zhichao, Yue Qiao, and Xihong Wu. Matching projection decoding method for ambisonics system. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 561–565, 04 2018. doi:10.1109/ICASSP.2018.8461515.

[31] Andrew Wabnitz, Nicolas Epain, and Craig T. Jin. A frequency-domain algorithm to upscale ambisonic sound scenes. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 385–388, 2012. doi:10.1109/ICASSP.2012.6287897.

[32] Leo McCormack and Archontis Politis. SPARTA & COMPASS: Real-Time Implementations of Linear and Parametric Spatial Audio Reproduction and Processing Methods. 2019 AES International Conference on Immersive and Interactive Audio;, 2019. ISBN 978-1-942220-27-5. URL http://urn.fi/URN:NBN:fi:aalto-201906203881.

[33] Leo McCormack, Archontis Politis, Ville Pulkki, et al. Parametric spatial audio effects based on the multi-directional decomposition of ambisonic sound scenes. In *23rd International Conference on Digital Audio Effects (DAFx2020)*, 2021.

[34] Leo McCormack, Archontis Politis, and Ville Pulkki. Rendering of source spread for arbitrary playback setups based on spatial covariance matching. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 371–375. IEEE, 2021.

[35] M. Kajala and M. Hamalainen. Filter-and-sum beamformer with adjustable filter characteristics. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 5, pages 2917–2920 vol.5, 2001. doi:10.1109/ICASSP.2001.940257.

[36] Brian G. Ferguson. Minimum variance distortionless response beamforming of acoustic array data. *The Journal of the Acoustical Society of America*, 104(2):947–954, aug 1998. ISSN 0001-4966. doi:10.1121/1.423311.

[37] J. Fernande, L. McCormack, P. Hyvärinen, A. Politis, and V. Pulkki. Enhancing binaural rendering of head-worn microphone arrays through the use of adaptive spatial covariance matching. *Journal of the Acoustic Society of America*, 151:2624–2635, April 2022. doi:10.1121/10.0010109.

[38] V. Pulkki, L. McCormack, and R. Gonzalez. Superhuman spatial hearing technology for ultrasonic frequencies. *Nature*, 11(11608), June 2021. doi:10.1038/s41598-021-90829-9.

[39] Archontis Politis, Juha Vilkamo, and Ville Pulkki. Sector-based parametric sound field reproduction in the spherical harmonic domain. *IEEE Journal of Selected Topics in Signal Processing*, 9(5):852–866, 2015.

[40] The Advanced Television Systems Committee. ATSC 3.0: Standards. URL https://www.atsc.org/atsc-documents/type/3-0-standards. Accessed: 2022-06-14.

[41] Dolby Laboratories. Dolby Encoding Engine User's Guide. February 2019. URL https://developer.dolby.com/globalassets/professional/product-manuals/dolbyencodingengineusersguide.pdf.

[42] Fraunhofer Institute for Integrated Circuits IIS. MPEG-H authoring suite. URL https://www.iis.fraunhofer.de/en/ff/amm/dl/software/mas.html. Accessed: 2022-06-14.