Immersive Audio Series - Part 3

Essential Guide

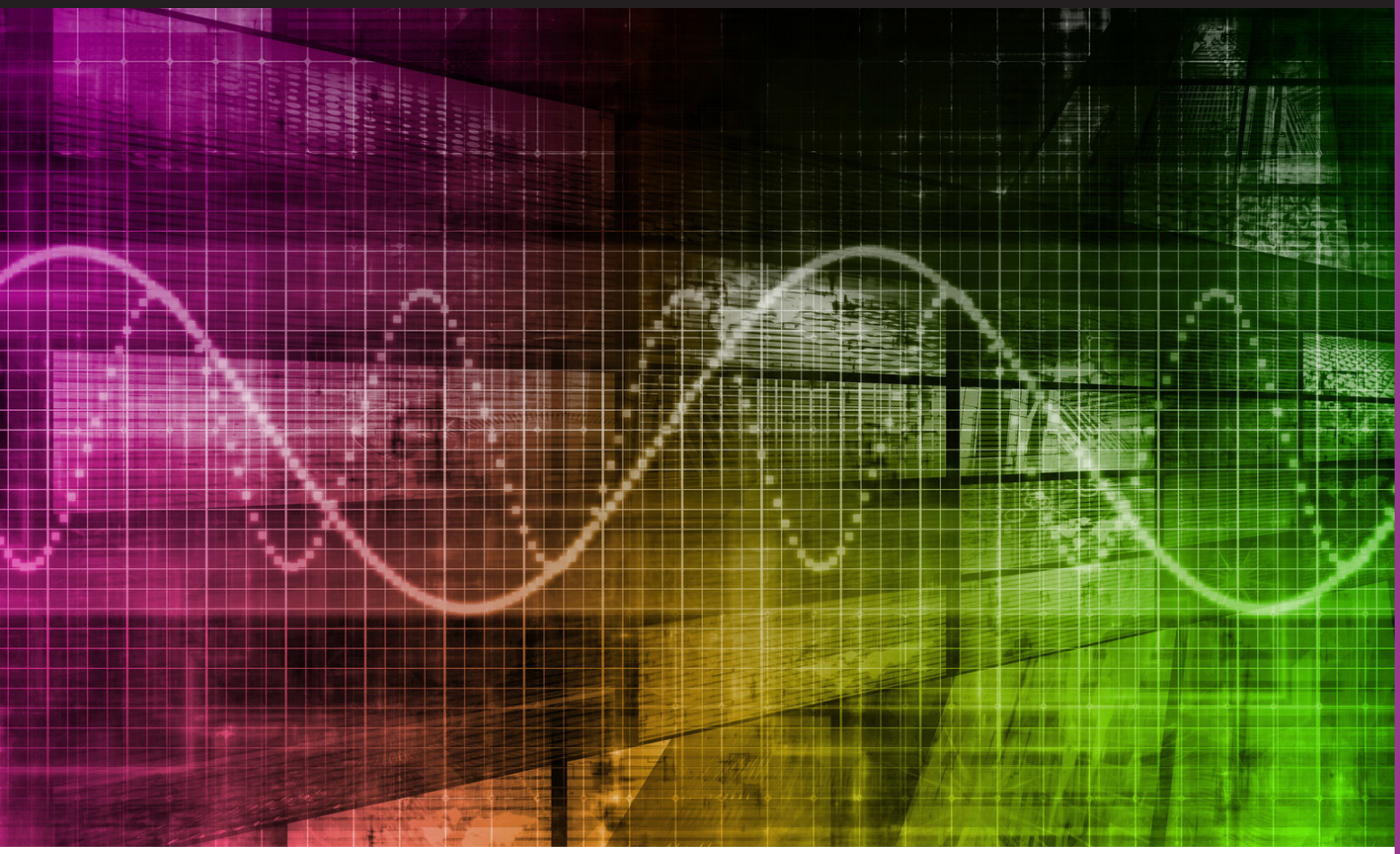
EG

ESSENTIAL GUIDES

Immersive Audio Essential Guide is a four-part series sponsored by;

Register today at thebroadcastbridge.com to receive your email publication notification for parts one, two and four.

Immersive Audio Series



By Paul Mac – Writer, Professional Broadcast Audio

Part 3 - Object Orientated Program

Paul Mac looks at the world of audio-based audio: Control options, choice, and some creative considerations.

The idea of an object in audio mixing shouldn't be too difficult to grapple with. It is, after all, just a channel - a channel that represents an object - anything that has a position in space. Under the hood, the mix in an object-based format is not the same as a mix in a channel-based format. In object-based audio the mixed channels are not printed into speaker channels at varying levels to give virtual positions; they are held back and stored with metadata that describes their position over time.

At the faders, though, the object -based world is exactly the same as in the channel-based world - channels still get panned; it's only the encoding and the decoding that changes. Although in immersive audio, the panning has an extra dimension…

How things find their positions in a three-dimensional world is an important practical part of production. In ambisonics and scene-based workflows that start out at ambisonic microphones, things are just exactly where they are.

Objects, however, need putting somewhere, which means you have to navigate a three-dimensional space.

The Dolby Atmos Production Suite is the standard set of plug-ins for mixing in Dolby Atmos with a DAW, which are then connected to the Dolby RMU for theatrical rendering. That comes with theatrical and VR workflow panners based on two different panning GUI models - XYZ and spherical. The differences between these two provide a useful basis to start getting your head used to three-dimensional panning and the different ways in which you can approach it.

XYZ panning - you could think of it as Cartesian coordinate-based - uses the X (left-right), Y (front-back), and Z (up-down) coordinates to manipulate position. Controlling three coordinates simultaneously but separately with the kind of finesse required in creative panning is not straight forward, but by linking, automating, and locking these three parameters in different ways, you can create some useful methods. This is where the Elevation Snap modes come in - a great example of assisted 3D panning. Here, the Z coordinate is calculated from the Y and/or X coordinates. In Ceiling Elevation mode, for example, you get an automatic lowering of the 'ceiling' (Z) in the front 20-percent of the room, moving towards the screen, so the Z coordinate is only dependent on the Y coordinate. Sphere and Wedge Elevation modes give you domed and peaked-shaped ceilings where Z is dependent on both X and Y.

In spherical panning - you could think of it as polar system - position is described by angle and distance (a vector). Consider a point in the middle of a space as the origin. That origin represents the position of the listener. Azimuth is the angle of an object from the listener in the horizontal plane, so at 0 degrees the object is directly in front. Inclination is the angle between the object position and the horizontal plane, so a 90-degree inclination puts the object directly above the listener and -90 degrees puts it directly below. Those two parameters help you navigate a sphere with a fixed radius, so in order to navigate the whole 3D space, you need control over distance - equivalent to changing the radius of that virtual sphere.

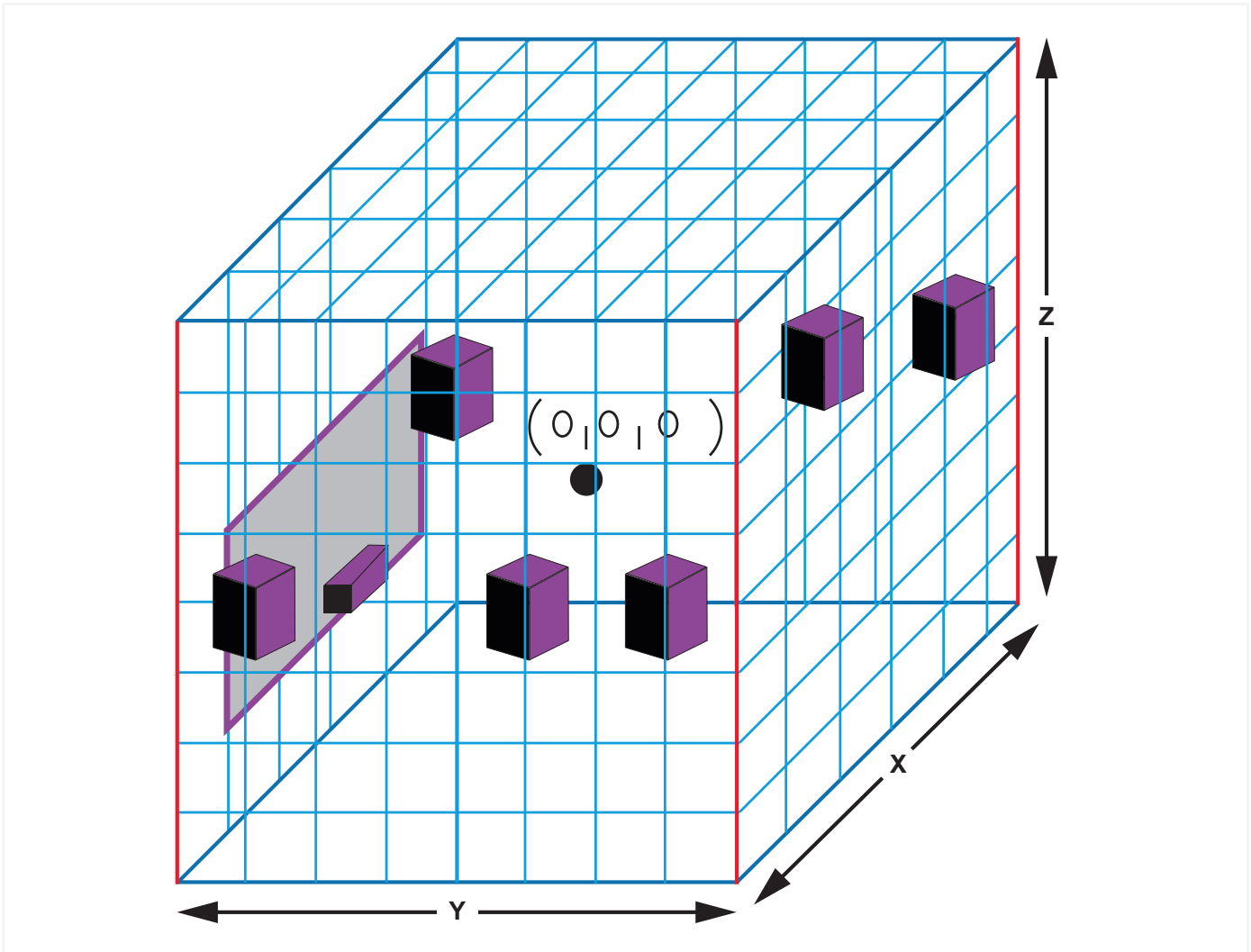


Diagram 1 – Cartesian XYZ coordinates used position object sound.

This spherical panning is particularly suited to VR workflows mostly because VR puts the viewer into that central head position, as opposed to traditional theatrical presentations where the visual aspect is static. Add in options for headtracking and the binaural distance model and it becomes a comprehensive VR model.

In any case, these two approaches are food for thought when it comes to the creative aspects of object-based mixing and how you might want to approach them and what suits the particular brand of immersive you're working on.

By the way, it's important to point out here that size matters too - the size of an object in Atmos basically defines decorrelation of an object from a point in space - its spread, if you will.

Are You An Object?

Another big decision in object-based mixing is, of course, what constitutes an object in your production. What sources you choose to be objects and what sources should be part of a channel-based bed, or part of a scene-based environment is partly a creative decision, but in many instances, it will be a practical one too.

In live broadcast production, few control rooms have enough hands for a dynamic object-based mix – though automated tracking technologies in sports might have something to say. Of course, just because they are not being moved doesn't mean they can't be objects - that's where static and selectable objects come in, like commentary channels and so on.

So, for live immersive broadcast the scene is an important tool - a speaker-independent representation of a three-dimensional space that can be rendered at the point of consumption to whatever listening environment is presented. Thus, each camera angle could have its own scene - or maybe not: Often having some kind of immersive environment - being part of the crowd - is enough for a viewer to take in the atmosphere without confusing switching of aural perspective. That is where the swell of ambisonic workflow might come in, whether it is native to the codec, or where maybe it involves decoding of ambisonics in objects or channel beds.

For production with the luxury of post-production, objects become a more dynamic consideration. The early 'big moments' included the flying fish scene in Life of Pi, which proved one of the big points when Atmos was first unleashed – the idea that mapping individual sources to specific speakers could make a massive difference to the intelligibility of those sources.

Your consideration might be practical, or creative or both, but extra planning is required if it's going to work.

The Gravity Benchmark

In 2013 probably one of the most iconic Dolby Atmos movies won Oscars for Sound Editing (Glen Freemantle) and Sound Mixing (Skip Lievsay, Niv Adiri, Christopher Benstead and Chris Munro). It's still a very special film and provides a real insight into the possibilities of object-based immersive audio when right-thinking, creative people get their hands on a new toy.

The film, directed by Alfonso Cuarón, follows a disastrous space shuttle mission that leaves two astronauts (Sandra Bullock and George Clooney) stranded in space. Freemantle called the film "an event... an immersive experience."

From the start, the film is an ultra-dynamic experience. The film goes from a serene space experience to full-on madness making full use of every speaker available and disorientating the viewer on purpose.

Freemantle used the subject matter and the creative objectives of the movie to come up with three main 'anchor' strategies for the project. First, while earth is not always visible, there are plenty of radio communications. Of course, inside a space helmet radio comms only come from your headphones, but in order to orientate the audience, the radio from earth is treated as an object and panned to wherever earth is currently positioned relative to the actors.

This relative panning is also employed for the actors and other elements. It's an unreal device that actually heightens reality. Genius.

Next they decided that even though sound does not transmit well through the vacuum of space, they would use the transmission of vibration through touch and into the air inside the space suit - creating a Foley library of muffled sounds that put you inside those space suits. That aspect establishes perspective.

Radio tuning effects are also used to emphasize the absence of communication and the potential for total abandonment. Even once you're aware of it, it's no less effective. And there is plenty of music in the film too – a channel bed that illustrates the action and joins with the other elements to provide an immersion that rises above simple object, scene, and channel considerations.

Gravity extensively used object sound and relative panning to achieve outstanding sound effects.

An early sequence in the film serves to illustrate this idea of all elements - touch, vibration, breathing, heartbeat, radio, and music – coming together. Debris hits the Hubble Space Telescope while astronauts Stone and Kowalsky are working on it. A massive collision is treated in silence as Stone is not attached to the colliding hardware – something Freemantle pointed out can be much more effective than even the biggest explosion: 'Sometimes silence with something huge is more frightening... especially if you're playing it off against other things.'

At the time, Niv Adiri summed up the commitment to this created reality and the immersive reward: "If you sell the concept early and it's a concept that works, you have to go for it. You can't say 'I won't do it on this cut' or 'I won't do it here'... just go for it. People get used to it and buy into it."

If you haven't already, watch Gravity, and be inspired.

The immersive audio toolkit is now many and varied – almost as many and varied as the platforms that exist to play it and the standards created to define it. The last two might be out of an engineer's hands to some extent, but the toolkit remains and can be used to create some amazing experiences.

Next time, we'll look at some workflows for implementing immersive audio in broadcast production.

SENNHEISER

Sennheiser

The Sponsors Perspective

Approach The Mic

Strategies for capturing immersive audio for scene and object-based audio.

dear VR.

For a truly immersive experience, cinematic virtual reality needs spatial sound - 360° spatial audio will make or break the immersive illusion. Just as with any video production, the key to success is correct recording. While traditional microphones still play an important role in virtual reality productions, they need to be augmented with spatial microphones that capture the full 360° ambience.

This is what the Sennheiser AMBEO VR Mic does - a single compact microphone that operates on the Ambisonics principle, allowing you to capture complete spherical audio from a single point in space via its four matched KE 14 condenser capsules in a tetrahedral arrangement. For playback, the audio is rendered binaurally, allowing you to virtually rotate the orientation of the perspective in all directions. Ambisonics is supported by all major post-production and playback tools on the market today. This makes Ambisonics the appropriate tool for Virtual Reality and all other applications involving immersive sound. Basically, you capture exactly what a listener would hear if he or she was standing in that position.

AMBEO on set.

Location Recording

Some care should be taken to record the Ambisonics signal correctly with regard to position and level, as certain errors cannot be corrected during post-production. A field recorder that has an AMBEO VR Mic mode such as the Zoom F8 will help to make this task easier. On location, the AMBEO VR Mic – fitted with the appropriate windshield or hairy cover – should be positioned as close as possible to the 360° camera as you need to patch the microphone out later in post-production, the same goes for the sound bag that accommodates the recorder. The VR Mic will usually be combined with additional conventional spot micrcophones such as wireless lavalier mics. This allows for increased flexibility during post-production, giving the mixing engineer greater control over the final experience.

Mixing

Mixing for cinematic virtual reality can be done in most standard DAWs as long as they support multichannel tracks, i.e. a minimum of four channels in a track. To support you in the mixing process, you should select an Ambisonics tool chain because most deliveries for cinematic virtual reality – including the AMBEO VR Mic recordings – are in Ambisonics. There are many tools to choose from, such as DearReality's dearVR tool chain or the free Spatial Audio Workstation available from Facebook. As all Ambisonics tool chains operate in B-format, you first need to convert the AMBEO VR Mic's raw 4-channel output signal to B-format using Sennheiser's free A-to-B format converter. The converter is available as free download for VST, AU and AAX format for your preferred Digital Audio Workstation for both PC and Mac. B-format is a W, X, Y, Z representation of the sound field around the microphone. W being the sum of all 4 capsules, whereas X, Y and Z are three virtual bi-directional microphone patterns representing front/back, left/right and up/down. Thus, any direction from the microphone can be auditioned by the listener during playback of Ambisonics B.

In mixing, use the recorded Ambisonics signal as the base ambience, then add signals from conventional microphones such as wireless mics and foley sound to emphasize and build your final mix. These additional conventional sound sources need to be spatialized so that they come from the correct point in space and match the video image and the ambience. This step is accomplished by your selected Ambisonics tool chain.

During mixing, it is important to monitor your Ambisonics mix. However, before you are able to listen to it, Ambisonics must be decoded. As cinematic virtual reality will in most cases be delivered over headphones, use a binaural renderer. Best practice is to monitor via the binaural renderer that is used on the platform or device that you will deliver your content to.

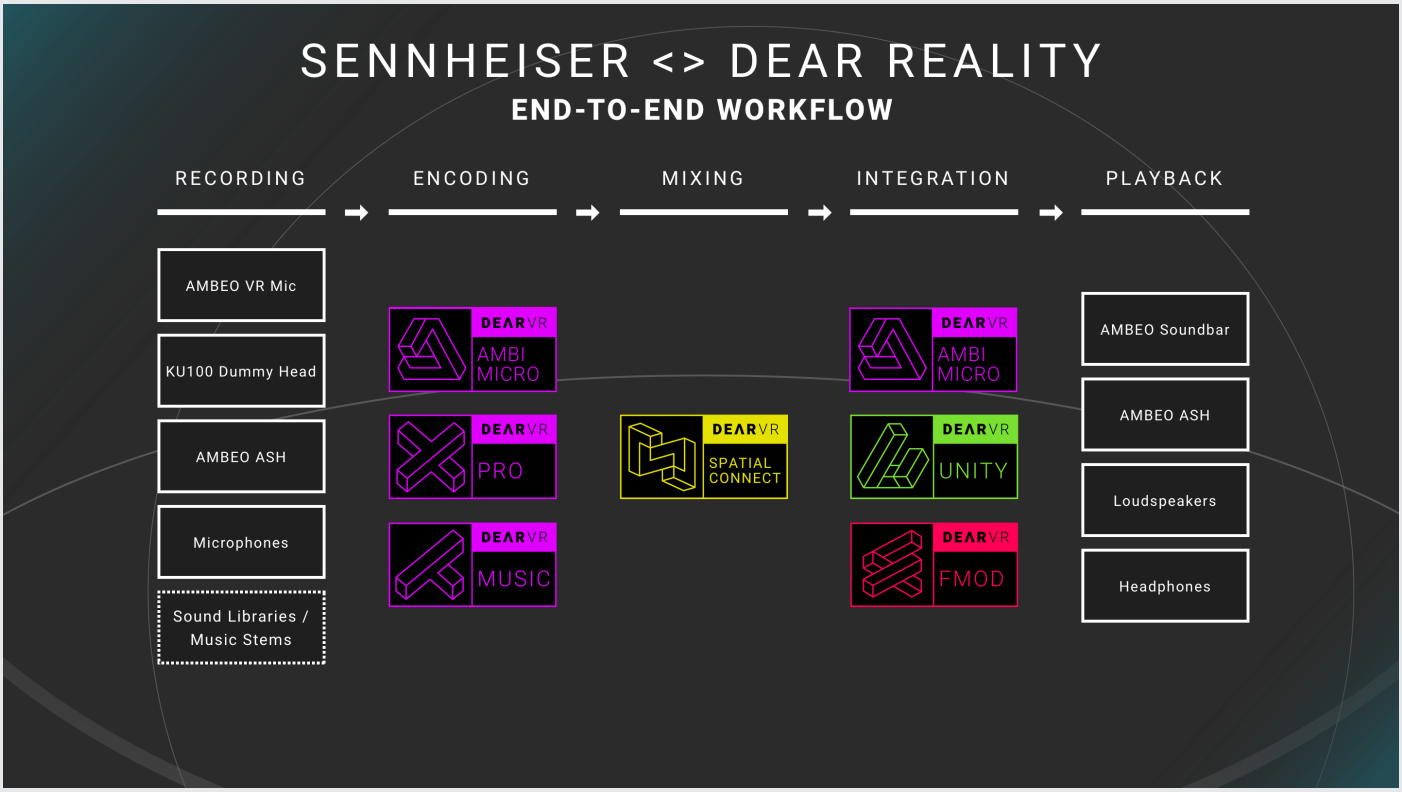
It should be noted that Ambisonics is the description of a sound field. Therefore, you should never work on any of the constituent tracks of an Ambisonics signal on its own. Always use Ambisonics mixing and editing tools if you want to modify an Ambisonics signal. When using a standard multichannel mixing tool, you must make sure that changes are applied equally to all four channels – otherwise you risk altering the spatial image of the Ambisonics signal.

Delivery

Ambisonics B-format is the audio format of choice for cinematic virtual reality. All major cinematic VR distribution platforms support Ambisonics B-format, including YouTube and Facebook.

To ensure that a file is viewed properly as a 360 video with 3D immersive audio, every platform requires its own metadata and has its own file format specifications. Please check the documentation of your targeted service. If you plan to distribute to your own app or custom platform, make sure to include support for Ambisonics decoding.

AMBEO on drumkit.



AMBEO end to end workflow.

AMBISONICS IN PRACTICE

RECORDING
You can use the AMBEO VR Mic to record full Ambisonics audio, or the AMBEO Smart Headset and the KU 100 dummy head to capture binaural audio. Simple mono microphone recordings or library sounds can also be used but require special encoding.

ENCODING
For further processing, the various input formats need to be converted to Ambisonics B-format and rendered binaurally for headphone monitoring. For these purposes, the dearVR AMBI MICRO includes the AMBEO A-to-B and AMBEO Ambisonics-to-binaural conversion libraries. Mono sources can be encoded to Ambisonics with dearVR PRO or dearVR MUSIC.

MIXING
dearVR SPATIAL CONNECT enables the user to mix virtual sound sources in VR and to control their position and levels in the dearVR PRO plug-in when in Ambisonics output mode.

INTEGRATION
By adding dearVR AMBI MICRO to the Ambisonics master bus in the DAW, the user can binaurally monitor the Ambisonics mix with headtracking using a VR headset. This makes for an easy assessment of the Ambisonics track on 360° video platforms or in game engines.

PLAYBACK
Use any pair of stereo headphones to monitor the binaural soundfield, or use the AMBEO Soundbar, AMBEO Smart Headset or loudspeakers for playback.

Genelec

The Sponsors Perspective

Rules Of Engagement

Genelec Senior Technologist Thomas Lund starts down the road to ideal monitoring for immersive audio by looking at what is real, and how that could or should be translated for the listener.

Natural listening is immersive. Whether in a room or outdoors, sound is all around us. Just two months old, we automatically recognize the direction of a sound, turning eyes towards a source, and half a year later we start using movement to change the perspective as an integral part of hearing and seeing when interpreting the world.

Actively reaching out using physical movement is a main element of human sensing, where the ground rules are laid before the age of two. Early in life we also already make use of individual and unique outer ears, to understand, for instance, what is direct sound and what is reflected sound. Like the imprint of a mother tongue, early sensory experience becomes a reference we're unable to ever fully escape, naturally rooted in our anatomy and the conditions found on planet Earth around the time of our childhood.

Such fundamentals are important to keep in mind when discussing multichannel delivery and reproduction, from personal binaural to immersive in-room systems. The essential question to ask is how well a given system is able to satisfy those basic human rules of engagement.

Early surround sound formats such as 4.2.4 and 5.1, had a limited ability to influence playback localisation and envelopment, while NHK's multilayer 22.2 sound system now incorporates both horizontal azimuth and elevation for a potentially much more natural listener experience.

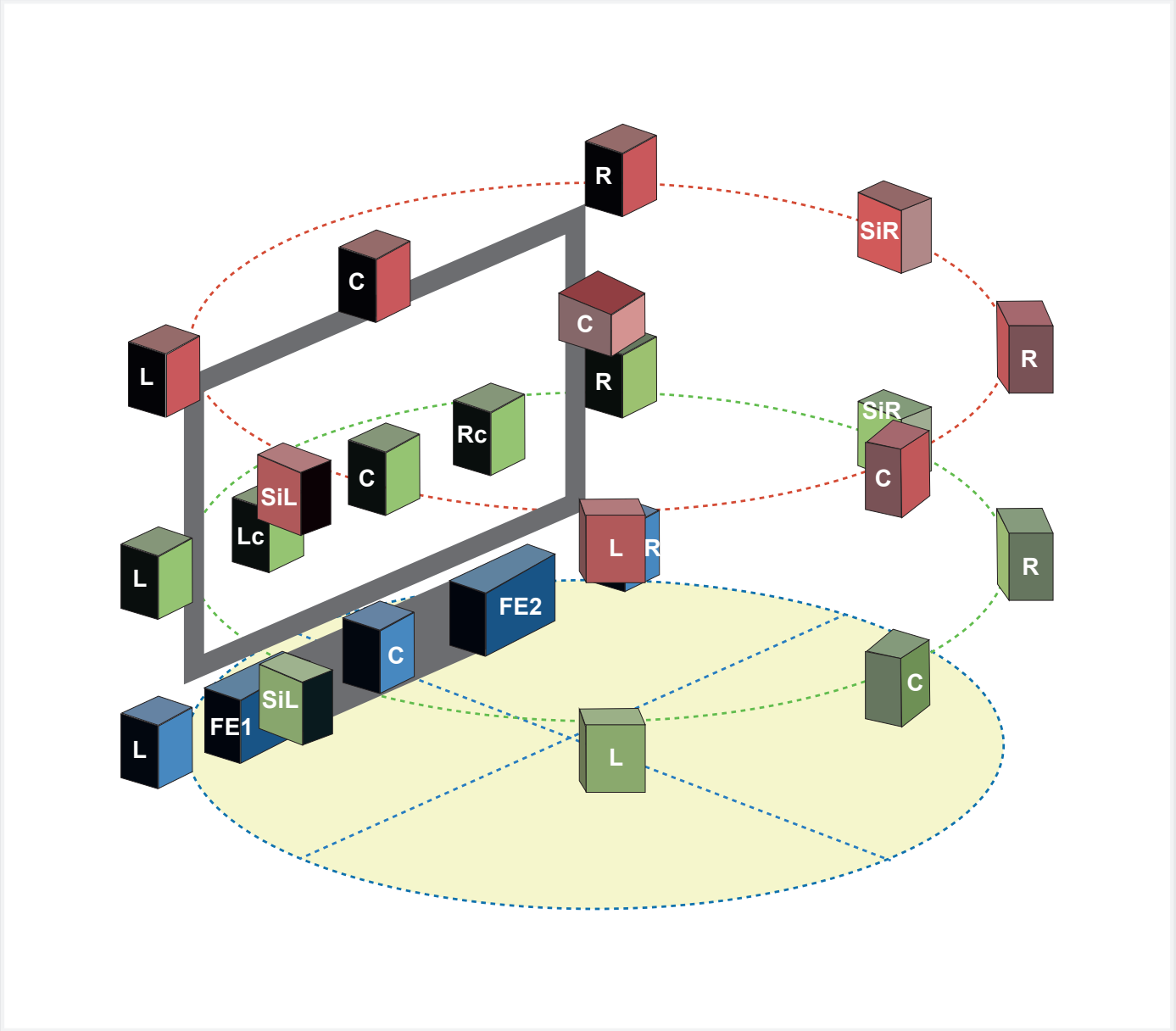
Before we get to the practicalities of immersive monitoring in part 4 of this series, let me provide an overview from multichannel audio research over the past decades.

From Surround To Immersive

When developing one of the first multichannel processors, TC System 6000, the limitations of 5.1 quickly became obvious, even with just one person seated at the ideal location. I was working at TC Electronic at the time, and we went on to develop a host of 5.1 and 7.1 algorithms, but magic didn't really start to happen until channels with elevation were added.

Genelec 8331 SAM loudspeaker

By 2010, the company studio had been retrofitted for 26 channel reproduction using active monitors, but I still remember the agony of adapting them to the room. Not only was it physically impossible to get the same distance to all monitors, each and every one sounded different, though they were the same type. An engaged engineer actually broke his arm climbing around to adjust switches on the back of those monitors to get their in-situ frequency response somewhat similar.



A layout for the surround sound component of Super Hi-Vision, developed by NHK Science & Technical Research Laboratories.

The shift from surround to more compelling immersive formats is reflected in ITU-R standards BS.775 and BS.2159. The two documents also describe the basic requirement that channels should all sound the same and be time- and level-aligned. However, they largely fail to explain the consequences of such requirements, leading many to believe monitors just need to be of the same type.

ATSC A/85 and EBU R128 provide more useful and practical guidance: “In-situ measurements of loudspeakers in control rooms, however, show strong deviations from the anechoic response of the loudspeakers, in particular due to room boundary loading conditions at very low frequencies, with standing wave modal effects through the range typically from about 80 Hz to 500 Hz.

For this reason, room equalization is highly desirable to the point of necessity for higher quality spaces.” In other words, each monitor must be frequency response corrected after its placement in the room, or you can’t trust what you hear.

New Immersive Studios

Immersive production is no longer reserved for theatrical content, but spreading into ambitious broadcast like NHK’s 22.2 format, OTT drama, enveloping music, and gaming. In any case, the listener will likely not be seated with 300 others, and a more personal reproduction may be assumed.

Production optimised for such scenarios should provide the professional with a more accurate and dedicated sweet spot, including him or her to making use of active sensing with head movements, thereby promoting content credibility and engagement. Delivery specifications also recommend monitoring at lower levels than for theatrical work, at 75 - 79 dB SPL, to ensure speech intelligibility and to reduce overall sound exposure.

New production requirements in turn mean new production possibilities, with less reliance on washed-out monitoring and more on conveying credible spatial contrasts and directional detail. Recently, precision monitors have become available that include compensation for placement and can be used at close range - for instance, the Genelec 8331.

Based on such technology, excellent immersive productions in 7.1.4 or 22.2 format may suddenly be realised in small rooms, between sidewalls 2m or even less apart.

Monitoring And Playback Using Headphones

The purpose of monitoring is to evaluate content in a neutral way, and to ensure good translation to other reproduction conditions. On-ear and in-ear headphones have not been ideal for this purpose, even just considering stereo production, because they exclude the influential external ear and movements from the equation. Headphones thereby break the link to natural listening described in the introduction.

Using generic headphones, important sources like human voice or tonal instruments are difficult to level, pan and equalise because mid-range frequencies translate randomly between people when using them. What you hear can be quite different from what the other person hears, even if you are passing the same set of headphones around.

However, more natural headphone playback is bound to become readily available soon, with several big consumer companies investing in egocentric “i”-solutions. The personalisation required for a landslide to happen has been demonstrated on the pro side, for instance with Genelec’s Aural ID method, so it’s only a question of time before large-scale personalisation for credible “iconsumption” becomes reality. Nevertheless, for a convincing experience there is no escaping the human rules of engagement: Consumer devices not only need to be statically personalised for direct sound from various numbers of directions, they also need to render in-room reflections personally; and to do it all coherently with head movements and low latency; like we experience in real rooms and everywhere else.

Though practical reproduction methods are lacking still, consumers will therefore shortly be able to enjoy immersive content better and more easily. Considering the production side, a standardised in-room monitoring system will often remain the option of choice because it translates perfectly to soundbars and other loudspeakers, as well as to ideally personalised headphone consumption.

Lawo

The Sponsors Perspective

Effectively Using The Power Of Immersive Audio

Lawo's Christian Scheck takes a tour of console functions and features that have a special place in immersive audio production, and how they are developing.

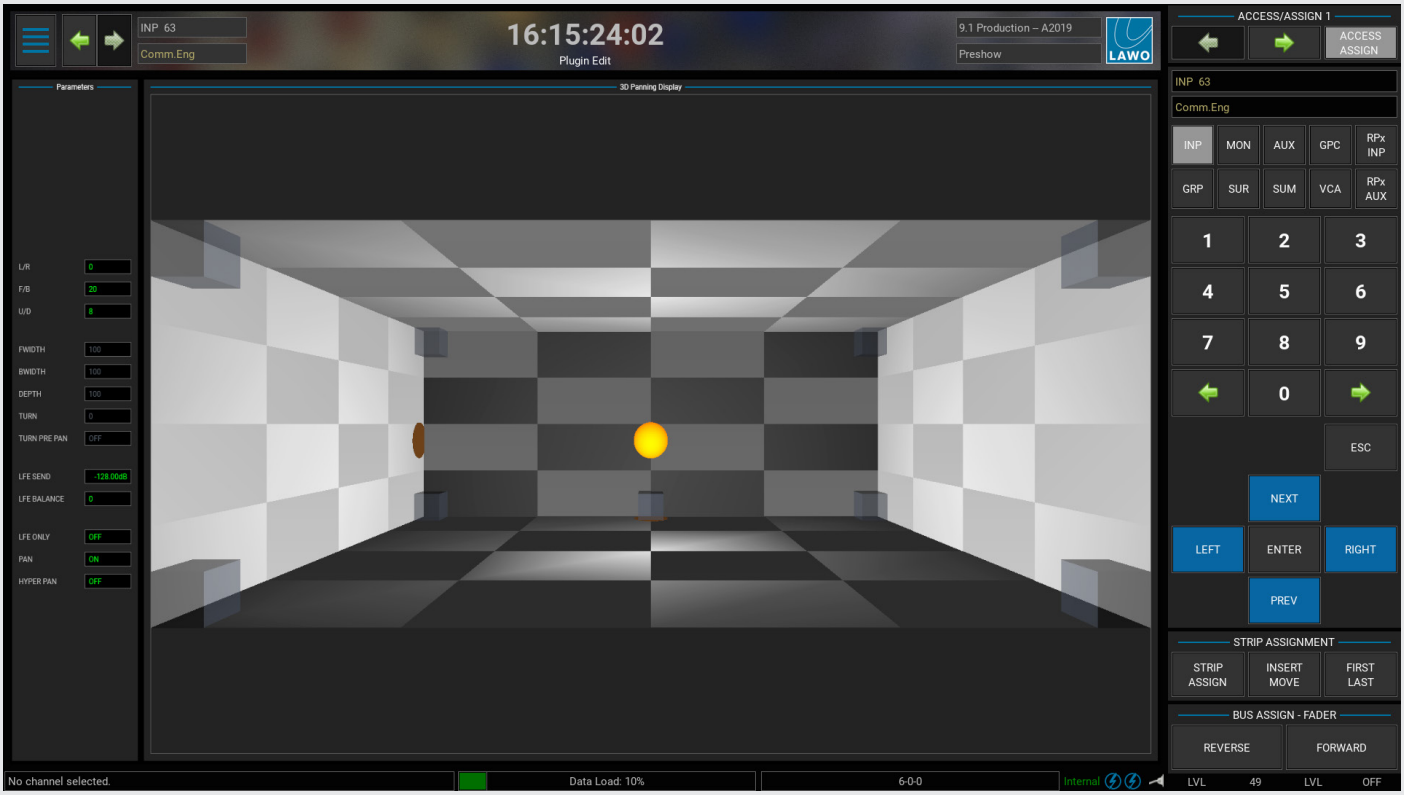
The LAWO mc²56 mkIII with immersive control shown on-screen.

One thing that is becoming increasingly clear regarding next-generation audio (NGA) is that, while a lot has already been accomplished, quite a few things still need to be done. Current achievements are the result of daring forays into uncharted territory where the only constant seems to be that Dolby's Atmos system, MPEG-H and a number of other 3D audio formats are here to stay and will play a prominent part in the evolution of immersive audio for the broadcast world.

Let us first look at the monitoring aspect of NGA productions. Since A1s can no longer predict the ways in which audio productions will be consumed by end-users, the only tools they have at their disposal are a growing number of presentations (delivery formats) they can monitor for consistency and quality.

The more presentations there are, the more you need to monitor at least occasionally. Long before anyone was aware of an object-based scenario, channel-based formats, like the momentous 22.2 immersive audio project developed by NHK in 2011, it was already clear that integrated monitoring would be a must-have feature. Luckily, the capability for this had been available on Lawo consoles for a long time. For today's NGA scenarios, A1s need all the help they can get—right at their fingertips.

The reason for this is simple: convenience. The more presentations audio engineers needs to monitor in a short time, the more important a convenient way of switching among them becomes. Moving around the control room to change settings on an outboard device (a Dolby DP590, say), is no option.



LAWO immersive in-console control.

A second consideration is what equipment should be available for monitoring in an OB truck or a control room. In specialized user communities, certain A1s have started to advocate the installation of soundbars in an OB truck, for instance, arguing that most NGA productions were likely to be consumed using these space-saving and relatively affordable speaker systems. The same applies to the binaural delivery format for headphones. Increasing one's chances to accurately predict the result requires monitoring the binaural stems as well.

User Interface

Being able to configure the authoring solution directly from the console is another way of saving time. Lawo's mc² consoles can retrieve the information from a Dolby DP590, for instance, and display it on their built-in screens.

This is enough in most scenarios as most broadcast productions use immersive sound in a fairly static way. Discussions about building a solution's full feature set right into a mixing console are still on-going.

Being able to remotely control the authoring tool from the console is, of course, very nice, but the console also needs to provide substantially more downmix firepower than most solutions offer today.

What would be an effective and reliable user interface? A system similar to the one jointly developed by NHK and Lawo, based on colored balls that provide a clear indication of a stream's placement in a three-dimensional space? Operators who have worked with it like the intuitive indication of signal placements.

Tools Rule

This leads us to the next consideration: panning signals in a three-dimensional space and the tools required to perform this effectively. Given the fairly static distribution of signals in an immersive broadcast production (hardly any signals need to whiz about listeners' heads), most operators seem to agree that a joystick for the X/Y axes and an encoder for the Z plane are easy to grasp.

Added flexibility is provided by functions like "X-Z Swap" on mc² consoles. It allows operators to assign the joystick to the X/Z axes and to use the encoder for controlling the Y axis.

So far, this has proven to be the right approach for live broadcast productions using immersive audio. Other controller types, however, are already under consideration.

One For All

It stands to reason that working with ten (5.1.4) or even more channels (7.1.4, 22.2, etc.) requires the ability to control all relevant busses using a single fader and that the metering system needs to accommodate a sufficient number of bar graphs.

Since nobody as yet knows for sure what the future will bring, the A__UHD Core architecture is prepared to support any multichannel format natively. This goes hand in hand with the required multi-channel dynamics, i.e. processors able to handle a multitude of channels (rather than a mere six).

All For One

In the light of the complexity facing audio engineers regarding mixing and—even more so—monitoring multiple presentations, multi-user operation looks likely to become the norm. Mixing systems will have to cater to such scenarios, with ample PFL/AFL support, CUT and DIM functions for all channels, and so on.

Going One Meta

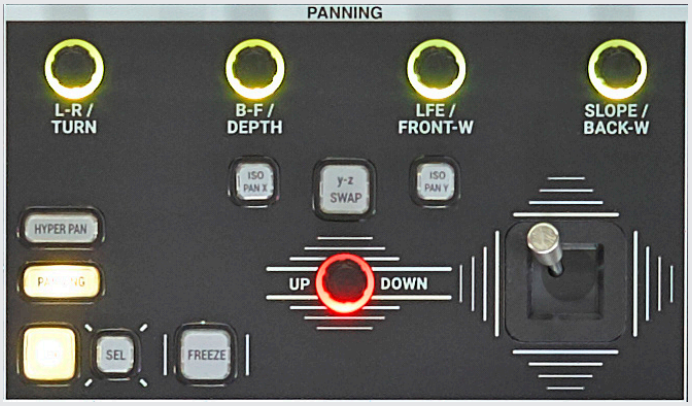
Next-generation audio goes beyond 3D immersive audio by providing tools that allow end users to personalize their listening experience. This is a major evolutionary step, which somewhat redefines the A1's paradigm.

While audio engineers have long lived in a comfort zone of at least a certain degree of objectivity regarding a "good" mix, personalization does away with this. Today's and tomorrow's A1s will at best be able to make educated guesses and monitor a rising number of presentations.

Still, this will likely be insufficient for satisfactory listening experiences in users' homes if the decoders stationed there have insufficient information regarding the stems they are receiving.

Enter the next buzzword for NGA productions: "metadata". That is, clear descriptions of the playback range of stems. Several approaches are being discussed regarding the kind of data that will be required by renderers to make sense of what comes in and what end users expect them to send out. A simple example in this respect is the commentary voice: with "low", "mid" and "high" settings, it will be easier for end users (and their renderers) to achieve a predictable result.

We definitely are living in interesting times…



LAWO MC²56 mkIII pan section.

WP
WHITE PAPERS

EG
ESSENTIAL GUIDES

MEDIA

CS
CASE STUDIES

Find Out More

For more information and access to white papers, case studies and essential guides please visit:

thebroadcastbridge.com

07/2019

Sponsored by Lawo, Genelec and Sennheiser