

Spatial audio

Channels, objects, or ambisonics?

Francis Rumsey
Consultant Technical Writer

Although channel-based audio is not dead in the current push for easily managed immersive audio workflows, alternatives such as object-based and scene-based audio have a lot to recommend them. This is because the latter offer the flexibility to handle the user interaction, context dependency, and multiformat aspects of modern systems. Papers from the recent AES conference dealing with these topics are summarized.

During the highly successful recent AES conference on “Spatial Reproduction—Aesthetics and Science” a substantial number of papers were presented on topics related to object-based and ambisonic treatment of spatial audio. In this article some of them will be summarized. Compared to the historical approach of treating spatial audio signals as “channel-based,” these more sophisticated approaches provide ways of representing a sound field in a way that can be rendered for a variety of different reproduction systems. Now that there are so many ways of reproducing spatial audio, both on headphones and loudspeakers, it makes sense to consider moving away from a conventional channel-based mixing and representation technique because otherwise one might need to make a different mix for every possible reproduction configuration.

The economics of sound production for movies, VR, and games, for example, may increasingly drive the industry in the direction of using either object-based or ambisonic representation of spatial and

immersive sound fields. It’s interesting that ambisonics was considered an esoteric and overly complicated concept for many years after it was first conceived. That’s probably because there wasn’t a sufficiently persuasive reason for people to consider adopting it in commonplace mixing and processing tools. These days the availability of so many formats, and the complexity of doing immersive audio in conventional channel-based ways, is leading to a resurgence of interest in flexible intermediate representation formats. The ease of processing signals using software and novel interfaces in digital audio workstations (DAWs) makes possible ways of handling spatial audio that would be difficult to implement on conventional mixing consoles.

BASICS OF SPATIAL AUDIO REPRESENTATION

By way of introduction to the papers discussed here, it might be helpful to clarify the basic differences between the ways of representing spatial sound fields using audio signals. Elements of each of these

approaches can often be combined successfully, as long as the system is capable of recognizing, handling, and rendering the different types of content (for example MPEG-H allows for all three methods of representation to be handled). Conventional channel-based audio (CBA, e.g., 5.1, 9.1) works by establishing amplitude and sometimes time differences between a specific number of loudspeaker channels for each element in the audio scene, including ambient sounds. Each channel corresponds to the signal fed to a loudspeaker that is located in a particular position. If the loudspeakers aren’t where the system expects them to be the spatial results will be unpredictable and maybe unsatisfactory. It works if you know how people are going to replay the signals, and there is a reasonable degree of identity between the layout used to mix the audio and the one used to replay it. When most audio reproduction used just two loudspeakers this worked quite well, but as more have been added and there are multiple immersive/surround formats the challenges have grown.

Object-based audio (OBA), in simple terms, relies on the idea that elements or objects within a sound scene can be represented by separate audio streams. Accompanying metadata are then used to describe where these elements should be panned to create the intended spatial scene. The content elements can be stored or transmitted separately, as numerous independent tracks or channels, along with this metadata. Some form of “rendering engine” is required at the reproduction end, to recreate the signals needed to realize the creator’s intention on some layout of loudspeakers or headphones. This approach lends itself to rendering for more than one possible layout of loudspeakers, and the rendering can be adapted depending on user interaction or context. It can get complicated if there are numerous elements in a scene, and the elements somehow have to be separated at the source end of the chain. The approach lends itself best to content that consists of separated monophonic source signals. The sonic result depends considerably on the efforts of the renderer, which is likely to be under the control of the ultimate consumer of the material. Hybrid approaches may involve representing conventional channel-based material as groups of audio objects that are assigned to specific loudspeaker locations (e.g., left and right channels of a stereo mix could each be audio objects that metadata say should be routed to appropriate loudspeakers).

Ambisonic-style representation (sometimes referred to in modern systems as scene-based audio or SBA) works by describing so-called spatial harmonics of the sound field. It does not represent the content elements of a scene directly, as does object-based audio, but rather it represents the sound pressure and velocity vectors of a spatial scene at a point or over a volume. Depending on the number of these spatial harmonics, the directional resolution of the scene can be higher or lower, and the results rendered to varying degrees of spatial resolution. More harmonics mean better resolution and the possibility to work

successfully over a range of listening positions. The great advantage of this format is that it is not loudspeaker-format-specific. The content can be processed by a rendering engine to suit a wide range of reproduction formats, both loudspeakers and headphones. Sound fields can be modified, rotated, turned upside down, and so forth, with ease. It suits sound scenes that are not easily taken apart into their constituent objects and that need to be represented as composite spatial audio scenes using a limited number of audio streams. One particularly useful application is for recording the spatial audio accompanying cinematic VR or 360 video, as we shall see later.

METADATA FOR OBJECT-BASED AUDIO

James Woodcock and a large number of colleagues from the UK discuss “A Framework for Intelligent Metadata Adaptation in Object-Based Audio.” As they point out, object-based audio “can be used to customize, personalize, and optimize audio reproduction depending on the specific listening scenario.” To that end they looked into a framework that used semantic metadata to describe the audio objects, loudspeakers, and room. An extensible software tool was developed for adapting those metadata in real time, incorporating the results from perceptual tests and/or perceptual metrics of, say, envelopment or speech intelligibility to enable optimum rendering of a sound

field. A block diagram of the framework is shown in Fig. 1. A Python software package known as the “Metadapter” was developed to implement some of these concepts.

One of the interesting points they make is that although object-based audio is said to be “format agnostic” (not depending on the reproduction layout), it is actually very rarely possible to create an identical listening experience on two different systems. That becomes immediately obvious when considering that objects intended to be behind or above a listener won’t be correctly rendered over two-channel stereo loudspeakers. This leads to the idea that one should somehow aim to adapt the rendering to maintain the producer’s intent as accurately as possible within the constraints of the system in question. When producing object-based audio sound mixers appear to adapt their approach depending on the semantic category of the object (e.g., whether it is dialog, foreground effects, background wash or music), say the authors.

A valuable use case was studied in order to test some of the ideas of this approach, looking at how experienced mix engineers downmix content for different formats. There is valuable expert information to be gained here about what a human engineer would do when faced with the task of preserving the rendering intent of a producer when using a less spatially sophisticated reproduction system than that used originally. It was found, for example, that changes in source position were used more often for foreground

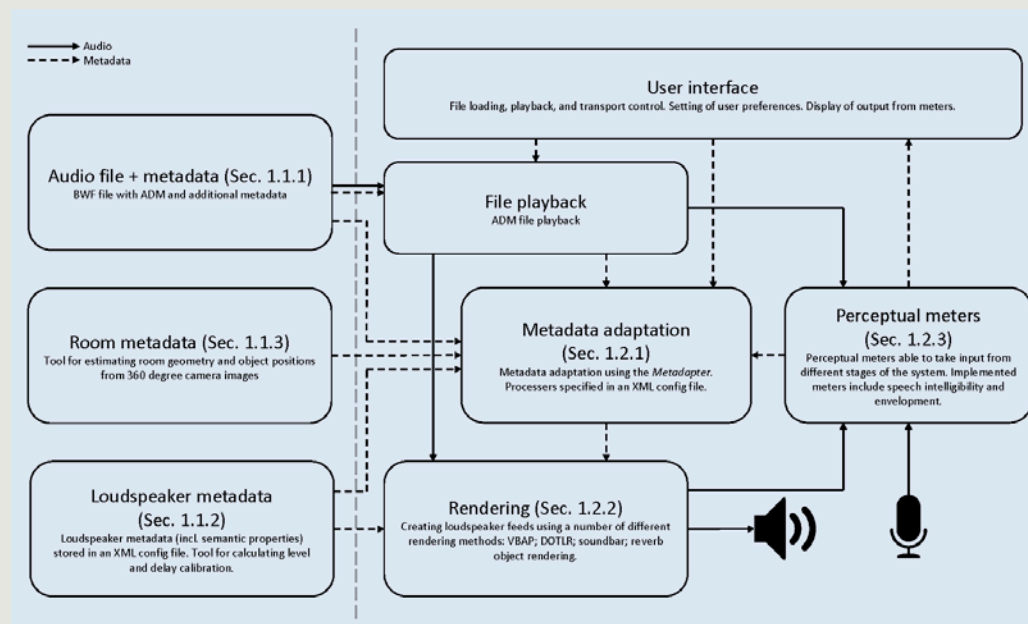


Fig. 1. Intelligent metadata adaptation framework (courtesy Woodcock et al.)

transient sounds, whereas changes in spatial spread were used more for continuous background sounds. Various ways of controlling the amount of envelopment were tried, based on either user preference or an aim to compensate for loss of envelopment arising from downmixing. Whereas currently the additional object- and scene-level metadata needed to facilitate these features are currently entered manually, then embedded in a BWF file, in future the intention is that they could be added by a producer using a digital audio workstation (DAW) or determined automatically from content. The various processors, perceptual meters, and rendering framework need to be integrated in such a way that they could be incorporated as plugins to work in a real production environment, it is said.

MASTERING AUDIO OBJECTS

Simon Hestermann, Mario Seideneck, and Christoph Sladeczek, in “An Approach for Mastering Audio Objects,” point out that although recording and mixing for object-based audio have been studied in recent years, there has been little attention to mastering. One could even ask whether the traditional concept of mastering is any longer relevant. The authors explain that mastering is the final step in conventional audio production, with the goal of improving the “overall sound of complete mixes without altering single elements of the production individually” (a definition apparently based on Bob Katz’s book *Mastering Audio*). Based on what they say later it’s clear that what is meant is that only the mixed output channels are modified in conventional mastering, rather than the mix elements. Reasons for it can include correcting mixes made in suboptimal monitoring conditions, or to match up tracks recorded at different times or in different conditions. It has typically involved dynamics and equalization processing and can also include making the results sound good when audio is encoded at low bit rates.

As shown in Fig. 2, an idea for the integration of object-based mastering into the audio processing chain is developed by Hestermann and his colleagues. In this conception the mastering functionalities are implemented prior to the loudspeaker renderer, but they say that mastering metadata could equally be sent to the renderer if enough processing power were available. Because of the added complexity and cost,

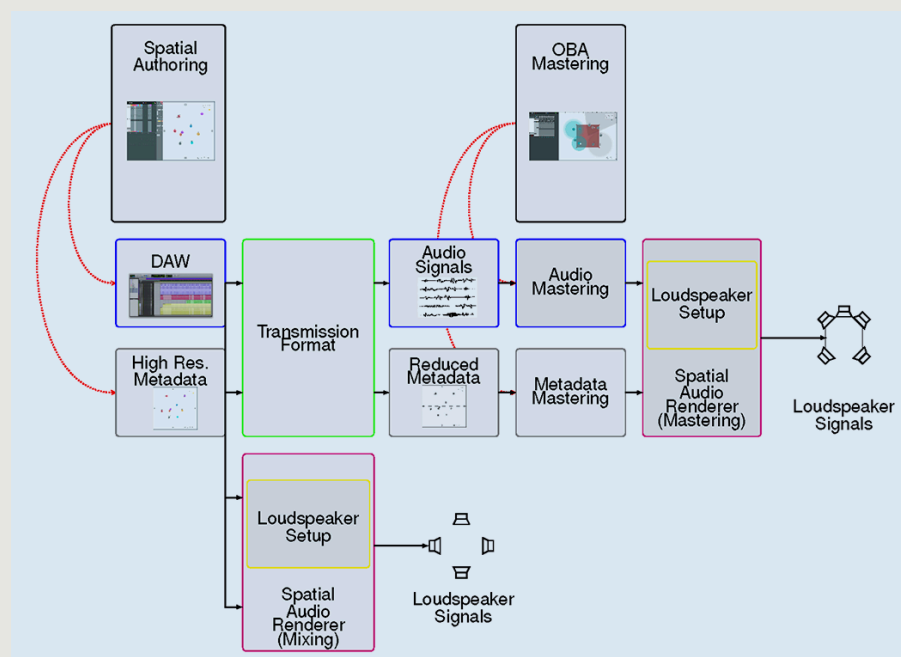


Fig. 2. Integration of object-based mastering into the audio processing chain (Figs. 2 and 3 courtesy Hestermann et al.)

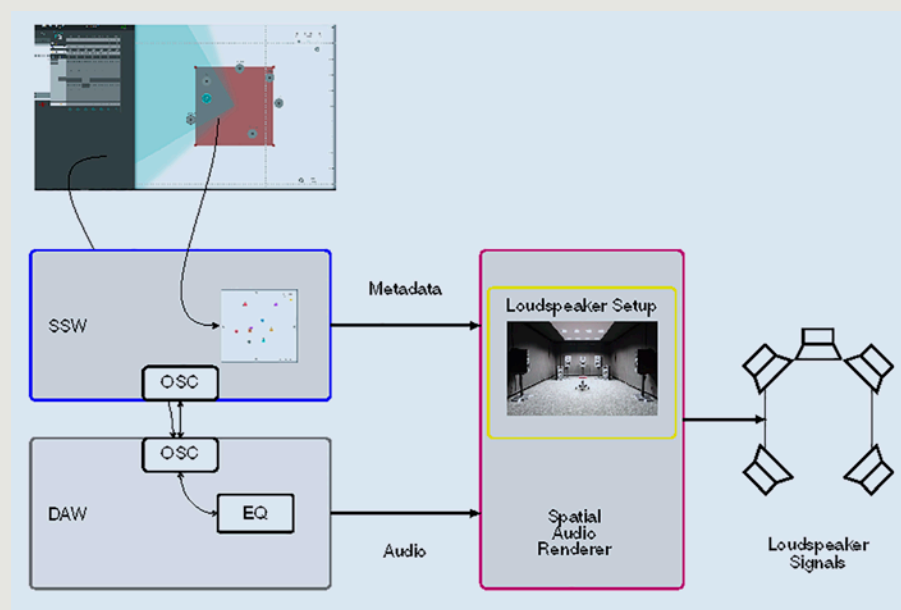


Fig. 3. Schematic of object mastering system architecture

though, they suggest that it’s unlikely to be possible for consumer devices in the long term. In order to preserve the idea of mastering as a form of high-level or overall adjustment of a sound product, the concept of “mastering objects” is introduced. These are supposed to alter multiple audio objects at the same time, up to and including the entire scene. They are to be thought of as containers that deal with a certain number of audio objects at once.

A prototype implementation of this approach was tried using the DAW known

as REAPER, because this employs an Open Sound Control (OSC) interface for exchanging metadata with other systems. Parameters from the object-based mastering interface are used to control specific audio streams in REAPER using OSC. The mastering interface itself was based on Fraunhofer’s “SpatialSoundWave” (SSW) technology, which employs a real-time renderer and a web server running applications to control the renderer. The way in which these are combined is shown in Fig. 3.

OBJECT-BASED POST-PRODUCTION

In the e-Brief “Creation of Immersive-Sound Content by Utilizing Sound Intensity Measurement. Part 3, Object-Based Post-Production Work,” Ritsuko Tsuchikura, Akiho Matsuo, and Masumi Takin discuss a new scheme for creating immersive sound content. It uses a form of full-sphere reverberation sampling known as VSVerb, combined with a 3D panning table in a mixing room that uses a Dolby Atmos object panner. Remarkably they appear to attempt simulation of the reflections sampled in a couple of studios by using 118 Atmos object tracks, one track per reflection, with each track controlled for the appropriate amplitude, panning, and delay of each reflection. This enabled the spatially accurate reproduction of short reverberation up to just under 90 ms after the direct sound, but the authors acknowledge that this is too short for the majority of practical purposes. The reflection parameters could be changed depending on sound source directivity or receiver position, and a split of the frequency range into LF, MF, and HF enabled spectral adjustment of the reverberation.

A 3D AV INSTALLATION BASED ON HIGHER-ORDER AMBISONICS

Shoken Kaneko and Hiraku Okumura describe “An Immersive 3D Audio-Visual Installation Based on Sound Field Rendering and Reproduction with Higher-Order Ambisonics.” The paper describes a case study of a 3D content-creation project done in conjunction with a sound artist. 108 loudspeakers and 6 subwoofers were installed at Yamaha’s Innovation Center in Japan, the configuration being shown in Fig. 4. Three projectors enabled a movie to be projected with a viewing angle of some 220° on the wraparound front wall. Using a Steinberg prototype 3D-enabled DAW, audio content of all three principal immersive audio types (CBA, OBA, and SBA) were imported. The OBA and CBA content was panned and encoded

to 3rd-order ambisonic (HOA-3) format, whereas the SBA content had been recorded using a 64-channel 3D microphone (ViReal Mic) and encoded to HOA-3. The resulting 3D mix could either be monitored using binaural rendering on headphones, or by decoding to an intermediate 42-channel CBA format and VBAP (vector-based amplitude panning) to the 108-channel loudspeaker array. The HOA decoder was implemented as a VST plugin, with it and the VBAP panner both implemented in the Max processing environment on a separate PC connected using a Dante network interface to the DAW. The overall plan is shown in Fig. 5.

One example of the flexibility of using HOA as an intermediate format for sound scene representation was that during initial production stages of the project the authors were able to use a 122-channel ViReal Dome for monitoring, whereas the actual installation used a very different layout and configuration with 108 channels. The 3D content was easily translated from one environment to the other by rendering the HOA content differently. HOA-3 uses just 16 channels to transmit the spatial scene, and that is all that the DAW currently offers, but the authors acknowledge that a higher order than

this would enable greater spatial resolution. If the DAW were to offer more than 3rd-order processing in the future, spatial source material could be re-rendered without needing to be remixed, in order to take advantage of any benefits of higher-order encoding.

RECORDING AND MIXING TECHNIQUES FOR AMBISONIC SOUND PRODUCTION

Tomoyuki Nishiyama and his colleagues from NHK discuss various ways in which they used ambisonic techniques to produce immersive audio for cinematic VR content in “Recording and Mixing Techniques for Ambisonic Sound Production.” Cinematic VR has pictures that cover a 360° visual

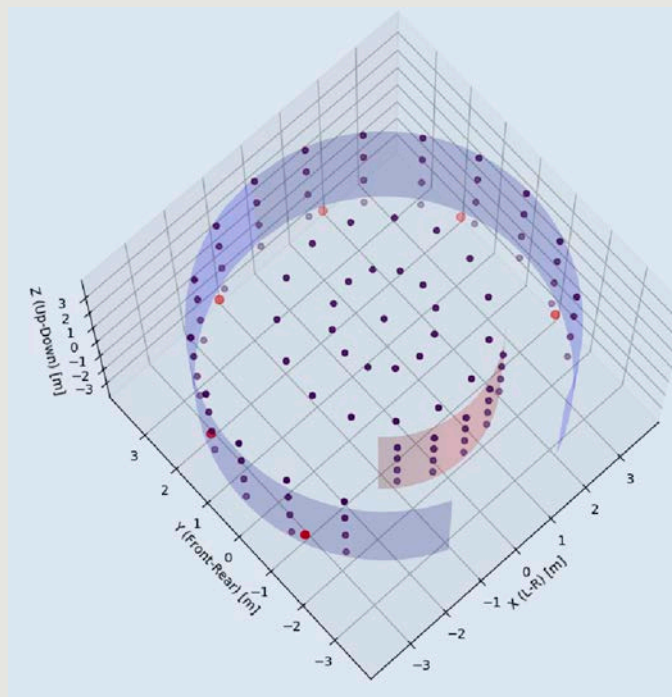


Fig. 4. Loudspeaker configuration of a large-scale 3D audio-visual installation. The blue and red arcs represent the front and rear wall, respectively. The purple and red dots represent the loudspeakers and subwoofers positions, respectively. (Figs. 4 and 5 courtesy Kaneko and Okumura)



Fig. 5. Overview of the audio system during the production phase of the large-scale 3D audio-visual installation shown in Fig. 4. The DAW and Max are running on two separate PCs, respectively. The purple and red dots represent the loudspeakers and subwoofers positions, respectively.

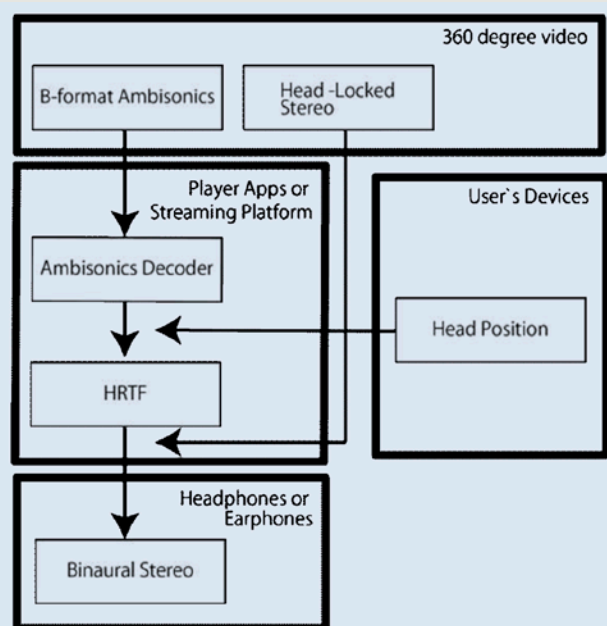


Fig. 6. Ambisonic production and rendering for cinematic VR (Figs. 6–8 courtesy Nishiyama et al.)



Fig. 7. Sennheiser's AMBEO VR B-format microphone mounted below a 360° video camera.

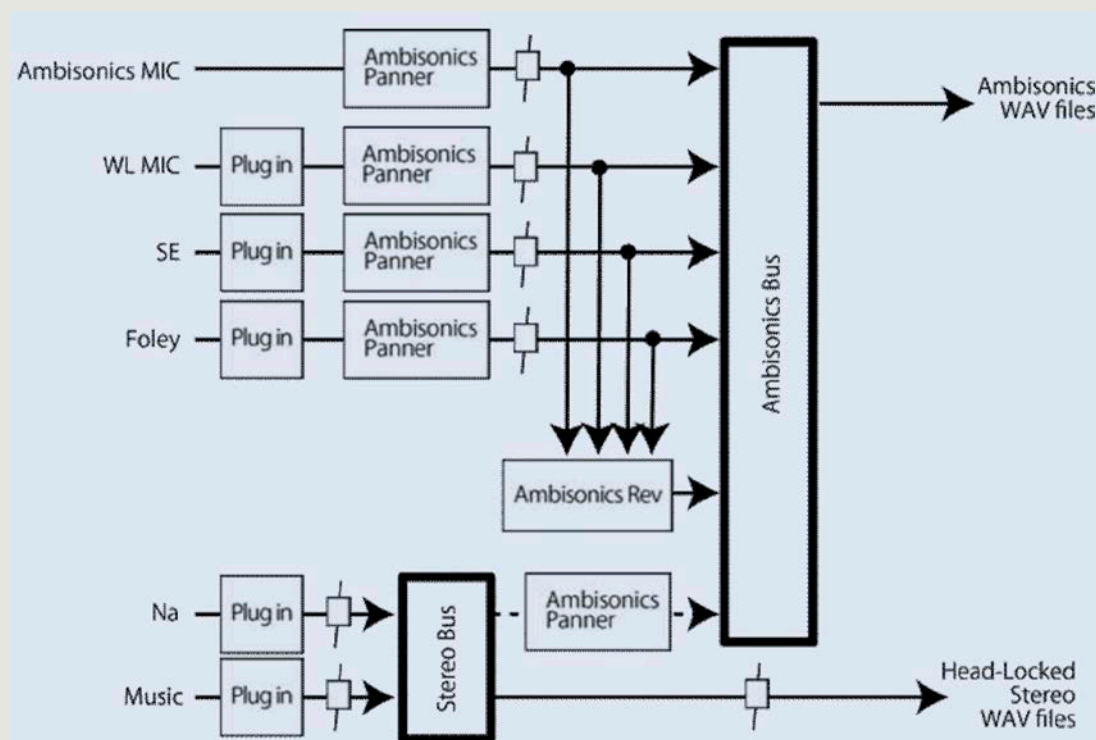


Fig. 8. Signal flow for ambisonic mixing

field, and viewers using head-mounted displays can turn around to see the scene from any angle. Viewers cannot, however, move around in the scene themselves, in other words the viewer is fixed in one physical position but can rotate their head realistically. (This is sometimes referred to as “three degrees of freedom” or 3DOF, whereas “six degrees of freedom” or 6DOF

allows full exploration of a virtual environment and requires not just head tracking but full positional tracking of the user, as well as an interactive virtual environment.)

Ambisonics would seem to lend itself rather well to cinematic VR or 360 video because it comes with the ready ability for spatial scenes to be rotated, while retaining all the relative positions of the sounds

within the scene, including the reverberant environment. Head tracking can be used to control the decoding of ambisonics to binaural HRTFs (head-related transfer functions) in order that the scene can be rendered according to the current rotation of the user's head, as shown in Fig. 6. As will be seen, some of the source content can be so-called “head-locked stereo,” which can be used for positionally steady sounds such as narration that don't need to be affected by head movements.

When recording live 3D content with a 360° camera, the authors used a small B-format (1st order) ambisonic microphone (Sennheiser's AMBEO VR, see Fig. 7) just below the 360 camera. One of the problems with recording such content is that 3D cameras see almost everywhere, so it can be quite hard to hide the microphones on a set. In this work lavalier microphones were also used on individual actors to capture their voices clearly, and the sound from these could be panned using an ambisonic panner in the DAW, as shown in Fig. 8. Mono sound effects and foley elements were employed rather than attempting to record them in an ambisonic format, because these too could be panned

appropriately into an ambisonic mix. With HOA panning, every track with an ambisonic panner took up nine (2nd order) or 16 (3rd order) channels, which meant that more than 500 audio signals were sometimes being handled by the DAW. The authors found that the number of tracks therefore had to be reduced in order to keep the DAW working properly.

The authors also discovered that peak management of the ambisonic signals was an important factor in avoiding clipping in the user's HMD. Because the decoding and binauralization parameters of a signal depend on the head rotation, signal overshoot can happen that was not anticipated during production. It was found that this needed checking under a variety of circumstances, and that a margin of some 10 dB below peak had to be left in the Ambisonic mix to avoid this. As a result certain very loud sounds such as explosions were consigned to the head-locked "conventional stereo" tracks so that they could be reproduced with sufficient audio level and without the risk that they would clip the user's headset audio. Compression of the ambisonic tracks had to be used a bit more aggressively than usual in order to get sufficient audio loudness for use in noisy environments.

CONCLUSION

Although channel-based audio is not dead in the current push for easily managed immersive audio workflows, alternatives such as object-based and scene-based audio have a

lot to recommend them. This is because the latter offer the flexibility to handle the user interaction, context dependency, and multi-format aspects of modern systems. Ideally, as we have seen above, some way of handling all three types of spatial representation in a workstation is likely to be needed.

OBA and SBA, though, rapidly proliferate the number of audio channels that need to be handled in a system. OBA may need to manage hundreds of separate elements for complex audio scenes, and HOA requires mixing buses with perhaps 9, 16, or more channels depending on the encoding order. Compared with the number of mix buses needed for handling 120 channels of CBA, though, this might seem reasonable, and you have the advantage that the content could easily be translated to a layout with a very different number of loudspeakers.

All this technical wizardry, though, is of relatively little importance if human perception and expertise is ignored. Any likeness between a sound field reproduced on a few loudspeakers and that heard on its creator's infinitely more advanced setup will need rendering that is informed by expert experi-

ence and perceptual quality metrics. There is still the interesting challenge of what to do with source content that was intended to be in one place and now seems to come out in entirely another because there are no loudspeakers in the right places, for example. Thankfully, at least in the consumer arena, a lot of this stuff is going to be auditioned on headphones where careful binauralization and head tracking may be able to resolve some of the problems. The number of situations where hundreds of loudspeakers are employed seems likely to be relatively small in comparison, limited perhaps to installations, movie theaters, and the like. Judging by the relatively modest take-up of true 5.1 audio systems (as opposed to sound bars) in people's homes, there seems unlikely to be a rapid move towards even more sophisticated multi-channel loudspeaker setups in these contexts.

Editor's note: the papers discussed in this article and others from the same conference can be downloaded from the AES E-Library at <http://www.aes.org/e-lib/>. AES members get free access to the E-Library.

FX100 Analog & Digital Audio Analyzer

Performance in your lab

Superior specifications cover numerous research and design applications over wide level and frequency ranges.

Speed in your line

System is optimized for measurement speed. Fast glide sweeps typically provide all relevant measurements on all channels in less than one second.



Modular system

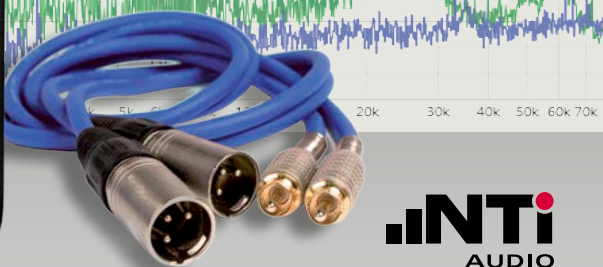
Configure to exactly meet your audio test needs. Add measurement channels, switchers, impedance measurement modules and interfaces.

FX-Control software

.NET-based FX-Control suite provides comprehensive, intuitive access to all controls and measurement functions. Control several instruments through one suite.

PureSound™

Leading-edge measurement technology reveals the speaker parameters, including Rub&Buzz, with a single stimulus. PureSound™ provides complete speaker characterization with unparalleled correlation to human hearing.



NTi Audio AG
Liechtenstein
+423 239 6060
info@nti-audio.com

NTi Americas Inc.
Portland, Oregon, USA
+1 503 684 7050
americas@nti-audio.com

NTi China
Suzhou, Beijing, Shenzhen
+86 512 6802 0075
china@nti-audio.com

NTi Japan
Tokyo, Japan
+81 3 3634 6110
japan@nti-audio.com

NTi Audio Korea
Seoul, Korea
+82 2 6404 4978
korea@nti-audio.com

www.nti-audio.com