Audio Engineering Society

Conference Paper

Presented at the Conference on
Audio for Virtual and Augmented Reality
2016 September 30–October 1, Los Angeles, CA, USA

Immersive Audio for VR

M. Altman[1], K. Krauss[2], J. Susal[1], N. Tsingos[1]

[1] *Dolby Laboratories Inc., 1275 Market Street, San Francisco, CA 94103-1410 USA*

[2] *Dolby Germany GmbH, Deutschherrnstrasse 15, Nuremberg, 90429, Germany*

ABSTRACT

It is commonly assumed that good VR audio content can be achieved by direct capture matching the camera perspective. However, crafting a compelling mix generally requires stepping beyond reality to offer an enhanced perspective on the action. The required artistic intent and creative integrity must be preserved from content creation to consumption by an end user, ensuring full immersion into the virtual world. In this paper, we review the different alternatives to delivering immersive audio for VR, contrasting physically-based and artistic-based approaches. We also offer an overview of audio content creation for VR, illustrating the benefits of object-based, artistically-driven approaches.

1 Introduction

Complete immersion in a virtual world requires 'tricking' one's brain into believing what they are sensing. Sight is limited by the field of view. Sound adds to what is not in view – a bull charging from behind, a rattlesnake on the right or even a whisper moving from the left ear behind the head to the right ear. Hence, content creators can leverage sound to direct the gaze of a user and effectively tell a story.

Object based sound creation, packaging and playback of content is now prevalent in the Cinema and Home Theatre [2], delivering immersive audio experiences. This has paved the way for Virtual Reality sound where precision of sound is necessary for complete immersion in a virtual world.

In VR, content creators need to be able to create object-based sound in a three dimensional space and encode their creations such that they are delivered, decoded and rendered binaurally (on headphones) and over speakers with precision and efficiency

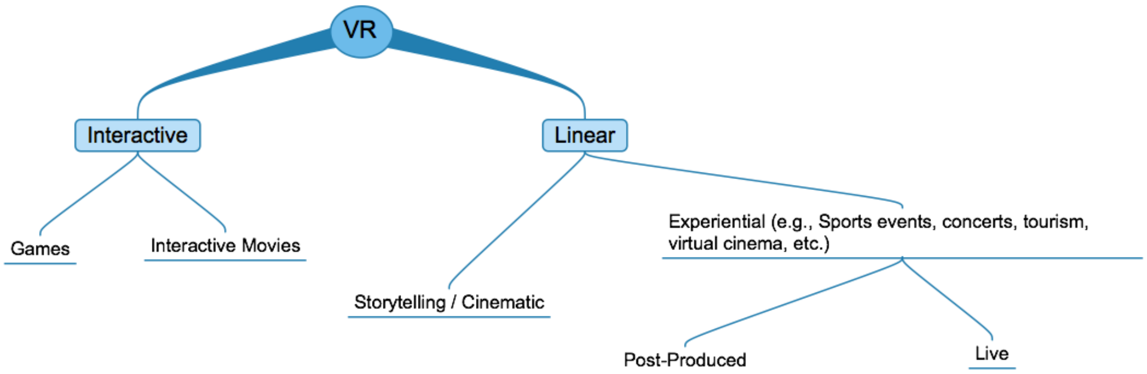

Figure 1. VR Landscape

when an end user listens to the content. It is commonly assumed that good VR audio content can be achieved by direct capture matching the camera perspective. However, crafting a compelling mix generally requires stepping beyond reality to offer an enhanced perspective on the action. The required artistic intent and creative integrity must be preserved from content creation to consumption by an end user, ensuring full immersion into the virtual world. In this paper, we review the different alternatives to delivering immersive audio for VR, contrasting physically-based and artistic-based approaches. We also offer an overview of audio content creation for VR, illustrating the benefits of object-based, artistically-driven approaches.

2 The VR Landscape

VR experiences can be broken down into Interactive and Linear use-cases (Figure 1).

Interactive VR includes Gaming and Interactive Movies, where the viewer is actively controlling the experience in real-time and is an active participant in the storyline. Linear VR includes Cinematic and Experiential events, where the viewer is able to control their viewpoint or select branch points in an overall linear timeline but not change the experience. The viewer may also be able to influence the audio story based on their gaze direction, for instance emphasizing some elements that are in their direct line-of-sight.

For gaming VR experiences, the immersive audio is generated by an audio middleware within the game engine. Game audio engines, such as WWise [3] or FMOD [4], render audio objects in real time to create the interactive audio. Generally, this type of experiences requires a download of audio and visual assets beforehand, as well as runtime logic to support the desired interactions.
Immersive audio for linear VR experiences is created offline in the case of cinematic content; it is produced live in the case of a live event. The audio can be encoded, packaged and streamed to the viewer, irrespective of the actual production process.

In both interactive and linear VR, the viewer is engaged with the content, able to look around the scene in all directions. With the advent of affordable motion trackers in VR systems, the viewer will soon be able to move around the scene, be able to peek around corners, walk around the space and change orientation to look at the scene from different angles with six degrees of freedom.

This freedom of movement and the ability to influence the audio experience at playback-time requires authoring and delivering content using a flexible audio format that enables spatial transformations and modification of elements within the audio mix.

In addition, a flexible audio format enables the VR content to be consumed across different playback environments. To date, immersive audio rendering for VR has largely meant rendering binaurally over headphones. However with immersive speaker playback systems now available in the home, an immersive audio experience (e.g. from a 360° video) can also be delivered by speakers in a living room shared by multiple listeners in the same room.

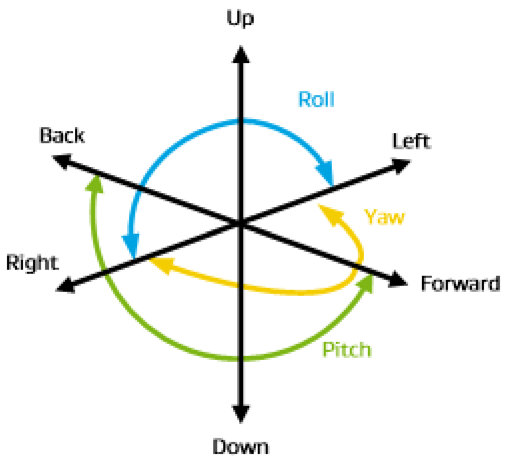

Figure 2. Six degrees of freedom (6DOF)

3 Flexible audio representations for VR

Two major classes of flexible audio representations are currently being used for VR applications: sound-field representations and object-based representations. Sound-field representations are physically-based approaches that encode the incident wavefront at the listener location. Well known approaches such as B-format or Higher-Order Ambisonics (HOA) represent the spatial wave front using spherical harmonics decomposition [7][8][14][15]. The spatial information of the sound field is directly encoded into the PCM waveforms of each harmonic signal, making these approaches similar to legacy channel-based audio representations. The signals can be further manipulated (e.g., rotated) and decoded over a variety of playback systems including binaurally over headphones. A variety of systems, such as A-format microphones or higher-order microphones (e.g., the Eigenmike) [11] can also be used to capture a soundfield representation. A disadvantage of HOA approaches is that they generally require a rapidly increasing number of PCM channels to encode more detailed spatial information; 4 channels for B-format/1st order but 16 channels for 3rd order. As a result they remain mostly practical at low orders and are mostly appropriate for 3DOF playback scenarios.

Alternatively, object-based approaches represent a complex auditory scene as a collection of singular elements comprising an audio waveform and associated parameters or metadata. The metadata embody the artistic intent by specifying the translation from the audio elements to the final reproduction system.

Sound objects generally use monophonic audio tracks that have been recorded or synthesized through a process of sound design. These sound elements can be further manipulated, e.g. in a digital audio workstation (DAW), so as to be positioned in a horizontal plane around the listener, or in more recent systems in full three-dimensional (3D) space using positional metadata. An audio object can therefore be thought of as a “track” in a DAW.
Similarly, interactive audio engines found in video games or simulators also manipulate sound objects - generally point source emitters - as the building blocks for complex dynamic soundscapes. In this case, they can incorporate very rich sets of metadata

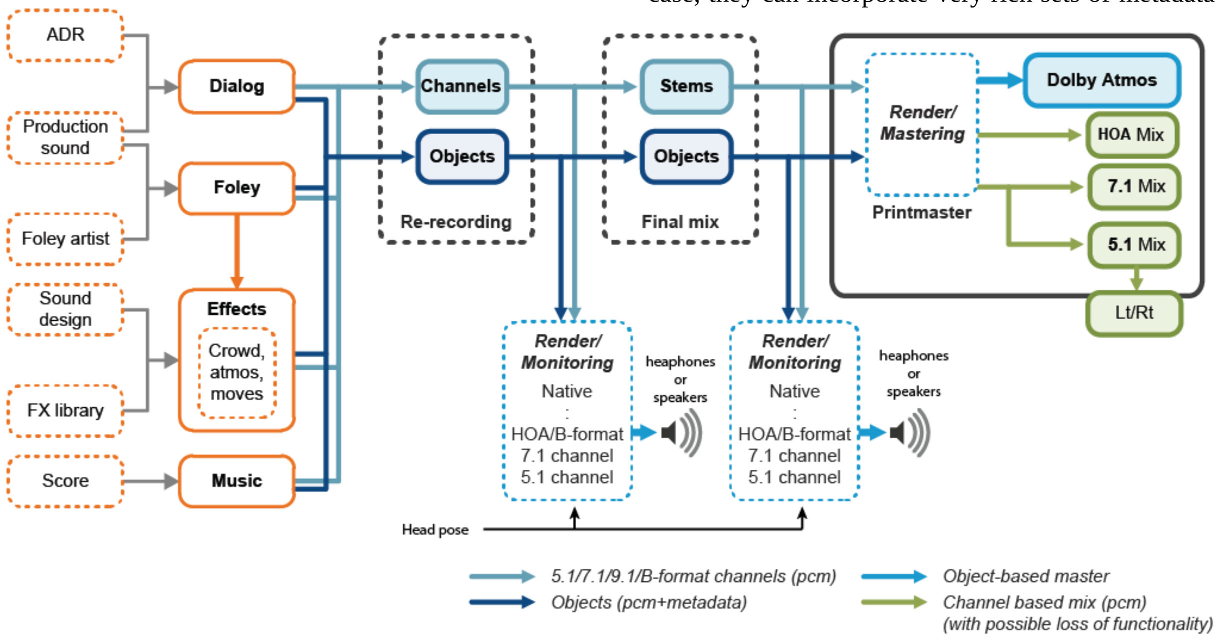

Figure 3. Authoring Workflow, showing a combination of Channels and Objects for theatrical or episodic production.

determining their behaviour.

Object-audio representations lend themselves well to the flexible processing required by VR applications since positional metadata can be dynamically modified, e.g. in response to head-tracking. Contrary to soundfield approaches they offer a better mix between spatial fidelity and interactivity as objects can be kept discrete and individually tagged with metadata indicating specific rendering behavior based on the desired application. Spatial accuracy is also only dependent on the metadata and therefore does not degrade with increased waveform coding or is not directly tied to the number of delivered channels. As a result objects are a good fit for applications requiring 6DOF playback with high spatial accuracy. However, they require an additional step in production workflows compared to pure channel-based solutions as metadata needs to be generated either automatically or by a mixer, which can be challenging particularly in live applications. Several solutions are available to mitigate these issues, for instance by converting spatial microphone signals to sets of objects or channels [5][10][11] as well as automated panning, e.g. based on video tracking.

4 VR audio production and post-production

The process of production and post-production for linear VR experiences is similar to traditional cinematic content. A set of audio elements obtained via spatial/soundfield recordings as well as spot microphones reach an audio mixing console or digital audio workstation where an audio engineer crafts an audio mix suitable for binaural reproduction over headphones. This creative process is paramount to deliver a high-quality, hyper-real experience. This experience goes beyond a single point audio capture collocated with the camera which can suffer from undesirable noise/elements and in general cannot lead to a satisfactory experience. Figure 3 shows an example of such a process. As the different elements get combined, the resulting mix can be monitored over headphones with head-tracking. The final result can be delivered over a variety of formats, either natively as objects, or alternatively pre-rendered into a soundfield re-presentation with the associated limitations in spatial fidelity and interactivity previously discussed.

4.1 Post-production for spherical and 6DOF content

An essential component of VR mixing is positioning the different audio elements of the mix in space (i.e. panning) so that they match the video reference. For traditional cinematic use-cases, this is achieved using a user interface similar to Figure 4, where the sound objects are positioned into a room in reference to a screen. For VR, the mixing interfaces must be adapted to account for the fact that the video/visuals can encompass the entire sphere or even an entire 3D region around the nominal listening position.

The left side of Figure 5 shows a VR specific panning interface adapted to spherical videos, where the panning space matches the equi-rectangular projection commonly used for video delivery. In this case, the interface controls azimuth / inclination / distance rather than the Cartesian framework used in Figure 4 on the right side. The egocentric frame of reference as shown in Figure 4 on the left side can be well suited to mixing content for 3DOF interaction but can quickly become too limited in cases where 6DOF interaction is possible.

For these more immersive use cases, using an allocentric/Cartesian frame of reference (Figure 6) where sounds are positioned relative to the environment is preferable. The right side of Figure 5 illustrates such a 6DOF VR mixing interface, where sounds can be directly positioned into the virtual environment.

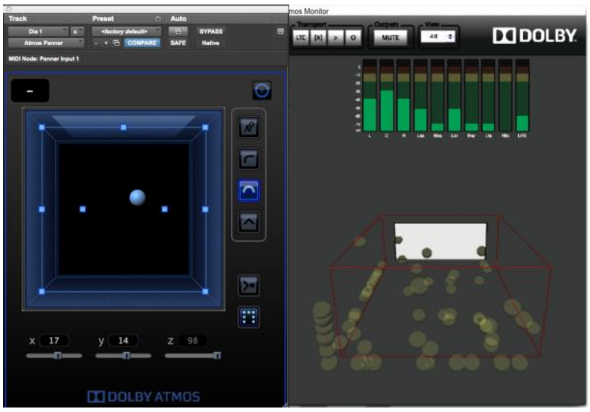

Figure 4.(left side) Dolby Atmos Panner plugin UI for ProTools™. This plugin allows mixers to specify spatial and rendering control metadata for objects. (right side) The Dolby Atmos rendering and mastering unit monitoring application showing rendered speaker feeds and objects

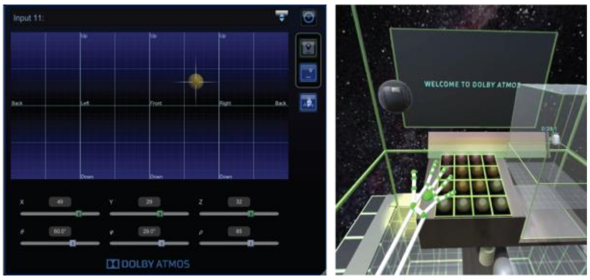

Figure 5. (left side) Extensions of panning user

interfaces to enable equi-rectangular panning. (right side) Full VR mixing interface with hand tracking and gesture-based interaction.

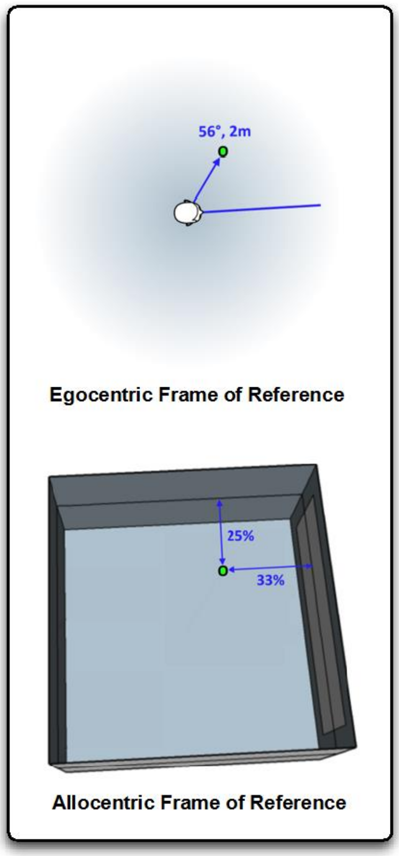

Figure 6. Frame of Reference for sound placement

4.2 Object metadata and controls for VR

Similar to traditional cinematic use cases [13], cinematic VR mixing often requires some audio elements to have specific playback-time behaviour. For instance, some non-diegetic background elements or music should preferably be kept 'head-referenced' i.e. non-head-tracked, while the diegetic sound effects or dialog should be 'scene-referenced' (i.e. head-tracked).

Similarly, it may be desirable for some elements to be rendered with higher timbral fidelity by bypassing binaural processing at playback time.

Another category of controls for VR applications determines the environmental model (reverberation, distance attenuation, source directivity, etc.). For spherical videos / 3DOF content, the environmental model is often pre-baked into the audio elements themselves. However, if 6DOF interaction is desired at playback time, these controls should be included as metadata in the delivered content so that the rendering algorithm can adjust the mix to the listener position.

Finally, a last category of VR specific controls relates to gaze-based interaction where the end-user can emphasize or even mute/unmute some of the elements in the mix by looking at specific points or directions.

These specific behaviours or properties can be easily authored and attached to the audio elements as object metadata. However, all these types of control are less compatible with a soundfield-type mixing and delivery, as multiple soundfield sub-mixes would have to be delivered individually, for each different playback-time behaviour. As a result, as these behaviours gain adoption from content creators, soundfield solutions may tend to increase bitrate and decoding complexity.

4.3 Live VR workflows

Similar to post-produced workflows [9], live workflows can be extended to support VR use cases. Figure 7 illustrates an example live VR workflow. A key component is the renderer/presentation manager that generates multiple mixes from a set of input audio elements, e.g. for the different camera viewpoints. This ability to output customized mixes for the different camera viewpoints is more critical for VR than for traditional broadcast use cases as a

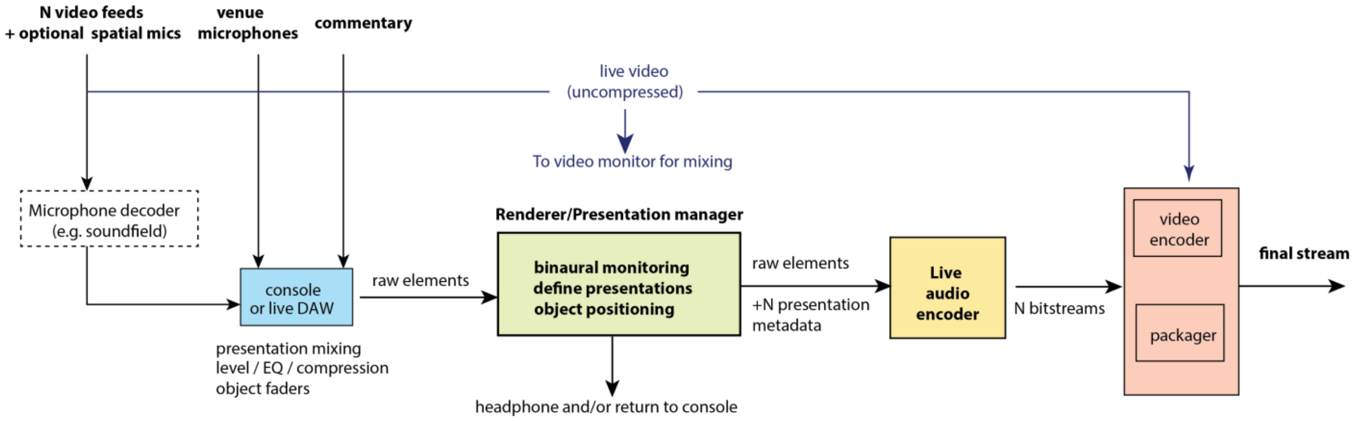

Figure 7. Example live VR workflow

tight audio-video consistency is a requirement for immersion. Object-based audio is well suited to create the hyper-realistic mixes required by professional live VR applications, where creating a soundscape that matches the camera viewpoint must be balanced with crafting an interesting mix. This may require enhancing far away elements that cannot be captured solely from the camera location but are nonetheless important to follow the action [1]. A solution is to author audio objects in 'world-space', i.e. the stadium or venue and let the presentation manager transform their position to match multiple viewpoints (e.g. for different cameras). In addition, an increasing number of systems are appearing which can be used to tie audio to real-time tracking systems in order to define dynamic audio objects such as the "Kick" system by LAWO (see also [6]).

4.4 Challenges of headphone mixing

So far traditional cinematic or live productions have often been authored focusing on traditional receivers like TVs or Set Top Boxes. Therefore the target loudness and dynamic range of the audio content has often been tailored for such receivers and playout via speakers. For consumption of audio on dedicated VR receivers like head mounted displays (HMD) with headphones this focus on just one dedicated receiver class must be broadened in order to cover different receiver classes. It is found that for rendering of immersive audio on headphones connected e.g. to mobile phones (HMDs) other, higher loudness target levels provide a better experience and immersion. Due to the fact that the levels of the rendered audio are then higher, the dynamic range of audio signals targeting such receivers also must be adjusted, to prevent the audio from clipping.

5 Use Cases to be considered for audio enabling immersive and authentic VR experiences

The following use cases benefit from immersive and authentic audio in VR.

1. Head tracking

A viewer enjoys a VR scene via his head mounted VR display (HMD) from a fixed point in the scene. He can freely turn and move his head to observe other details in the scene and may follow the actors by listening and watching. He likes to change his view to explore the details in the scene, looking around freely. At the same time the user wants to have an accurate binaural representation of the scene, so that he can follow the story without the necessity to (visually) search the virtual scene to find the main action. Accurate audio rendering supports the natural behavior of people to look around them, and then are able to quickly re-focus on a sudden event by 'hearing' accurately where it happens.

2. Immersive, social, simultaneous viewing

The setup of this use case is identical to the previous head tracking use case. The difference here is that the video is consumed on an HMD, but the audio bit stream is cast to an AV-Receiver. Instead of being rendered on headphones the AVR renders the audio to attached speakers. It is the same audio bit stream as in the previous use case, but sent to the AVR instead of being rendered on headphones. This enables social interaction of users in simultaneous shared VR spaces (a living room) with a deeper level of immersion.

3. VR multicast to many VR receivers

A broadcaster delivers a version of its new Reality-TV format as a 360° VR program to his viewers for consumption via VR equipment. The video and audio of the VR program is delivered via multicast to many viewers simultaneously. Each viewer is enabled to freely walk around the virtual environment and watch the program from different positions. The experience/sound of the scene adapts continuously as the viewer moves and always provides the viewer with an optimal experience, a personalized rendering of the audio scene, according to the individual viewer's virtual position in the scene.

4. Addressing broadcast receivers and VR receivers without simulcasting

A program provider wants to transmit a new gameshow as a 360° VR program over an IP network to its viewers for consumption via VR equipment, but at the same time also to its viewers using a regular TV to view the program. Due to limited resources the network operator wants to deliver just one and the same stream to traditional TV receivers and VR-enabled receivers. Audio rendering via the TV receiver follows the direction of the viewport provided. Audio rendering on the VR equipment follows the head-tracked position in the scene and viewing direction of the viewer.
The TVs display the main view from the stream following the main action (viewport). At the same time VR receivers would enable the full 360° view for users that experience the program in VR.

5. Camera on Racing car (on board 360° camera)

The racing sports league mounts omnidirectional cameras on every car in a race, in order to allow spectators with VR equipment to follow the action from the camera's viewpoint on top of the racing cars. As the racing cars drive at high speed they pass slower cars and static objects along the racing track. The sounds emitted from objects along the track (e.g. the horns of an Ambulance, or the engine sound of another car which is picked up directly by a microphone next to its engine) are accurately rendered (incl. the Doppler Effect) according to their position relative to the moving car.

6. Personalization and Accessibility

A content author wants to create cinematic immersive VR experiences reaching as many viewers as possible. The dialogue in the cinematic content therefore needs to be translated into different languages in order to also reach an international audience. In-line with state of the art film productions, the cinematic content undergoes a sophisticated post-production process with separate recording of Dialogue, Music, Foley and other effects. The VR consumer can select from a multitude of available languages in the audio track and can also change the dialogue loudness of individual actors in order to better follow the dialogue.

7. Augmented reality or Mixed reality

A tourist walks around sight-seeing in a foreign city where he has never been before. On his AR display the user gets visual information from an avatar/ virtual narrator about the sights worth visiting. While he is walking by the site the narrator explains historic. As the tourist approaches a sight, the narrator takes a position right next to the site. The tourist still needs to be able to pay attention to the busy environment while listening to the narrator. He can turn his head, move closer or further away from the narrator, interact with the real environment, e.g. move aside as somebody passes by, while the position of the speaking narrator remains where it is.

8. Diegetic – non diegetic rendering

In a VR documentary a viewer can experience rare and almost extinct wildlife through VR. An off-screen narrator explains the behavior and other details of the wildlife. The viewer can freely move his head and always experiences the sounds made by an animal from a dedicated place in the scene. As the narrator describing the scene is not visually present at a dedicated position in the scene, any head movement of the viewer does not lead to any difference in the rendering of the narrator's voice. This is done as an artistic element to even more clearly create a perceptual differentiation between the sound of the wildlife and the voice of the narrator.

9. Partial binaural rendering

A music fan is watching a live concert of his favorite band via a VR head mounted display and headphones. He is watching the scene from a camera position that is right on stage. He is experiencing the show in VR as if he is a member of the band. The musicians' instruments are picked up by microphones and are mixed by a professional mixer for the audience. The sound of some instruments e.g. a snare drum usually get compromised, i.e. loose attack and precision when they undergo an HRTF based binauralization process for spatial rendering on headphones. Therefore the audio from these instruments is marked specially to be omitted in the binauralization process. Those sound sources will be rendered unmodified, without application of binaural HRTF cues.

6 Conclusions

Advancements in spatial audio content creation, distribution and delivery are now capable of bringing more lifelike, scalable and interactive audio experiences to consumers across a wide range of devices and applications, including VR.
Object based audio perfectly complements the visual experience of Virtual Reality and is key to leveraging the overall experience to become truly immersive.

Object Audio:

- Natively enables complete movement along 6 degrees of freedom for the viewer in the virtual space, including translational movement.
- Allows for high spatial resolution due to objects rendered in a scene according to their positions, rather than in a mix.
- Provides the required flexibility, extensibility and precision to enhance the overall VR experience significantly.

Object Audio has already been proven to be viable and essential to improving immersive experiences in more traditional forms of entertainment like Cinema and Broadcast, and we have found that Object Audio for sound reproduction in VR as well provides a fully immersive experience. For the use cases mentioned in this document Object Audio is essential.

7 References

[1] J. Riedmiller, N. Tsingos, and S. Mehta, P. Boon "Immersive & Personalized Audio: A Practical System for Enabling Interchange, Distribution & Delivery of Next Generation Audio Experiences", SMPTE 2014 Annual Technical Conference & Exhibition

[2] C. Robinson, N. Tsingos, and S. Mehta, "Scalable format and tools to extend the possibilities of cinema audio," SMPTE Motion Imaging Journal, Nov. 2012. [11] F. Rumsey, Spatial audio. Taylor & Francis US, 2001.

[3] "WWise" [Online] Available: https://www.audiokinetic.com/products/wwise/

[4] "FMOD" [Online] Available: http://www.fmod.org/

[5] XY-stereo Capture and Up-conversion for Virtual Reality. Nicolas Tsingos, Pradeep Govindaraju, Cong Zhou, and Abhay Nadkarni. AES Conference on Audio for Virtual and Augmented Reality 2016 Sept 30 – Oct 1, Los Angeles, CA, USA

[6] Giulio Cengarle, Toni Mateos, Natanael Olaiz, and Pau Arumi. A new technology for the assisted mixing of sport events: Application to live football broadcasting. In Audio Engineering Society Convention 128, May 2010.

[7] R. K. Furness. Ambisonics – an overview. In AES 8th International Conference, Washington, D.C., 1990.

[8] D.G. Malham and A. Myatt. 3D sound spatialization using ambisonic techniques. Computer Music Journal, 19(4):58–70, 1995.

[9] Mark Mann, Anthony W.P. Churnside, Andrew Bonney, and Frank Melchior. Object-based audio applied to football broadcasts. In Proceedings of the 2013 ACM International Workshop on Immersive Media Experiences, ImmersiveMe '13, pages 13–16, New York, NY, USA, 2013. ACM.

[10] J. Merimaa. Applications of a 3D microphone array. 112th AES convention, preprint 5501, May 2002.

[11] J. Meyer and G.W. Elko. Spherical microphone arrays for 3D sound recording, chap. 2 in Audio Signal Processing for next-generation multimedia communication systems, Eds. Yiteng (Arden) Huang and Jacob Benesty, Bosten, Kluwer Academic Publisher. 2004.

[12] V. Pulkki. Directional audio coding in spatial sound reproduction and stereo upmixing. Proc. of the AES 28th Int. Conf, Pitea, Sweden, June 2006.

[13] Charles Q Robinson and Nicolas Tsingos. Cinematic sound scene description and rendering control. In Annual Technical Conference Exhibition, SMPTE 2014, pages 1–14, Oct 2014.

[14] Evolving views on HOA: From technological to pragmatic concerns. Jérôme Daniel. Ambisonics Symposium, June 25-27 2009, Graz, Austria.

[15] Further study of sound field coding with higher order ambisonics. Jerome Daniel and Sebastien Moreau. 116th AES convention. May 8-11 2004, Berlin, Germany.