PRODUCING 3-D AUDIO

JUSTIN PATERSON AND GARETH LLEWELLYN

Abstract (REVISED)

Arguably the most rapidly expanding market for audio production is that of 3-D audio – 360° spatial audio with representation of height. Such playback is becoming commonplace in cinemas, and multi-speaker home setups are following. As these mature, greater convenience and enhanced functionality will likely increase domestic uptake.

In parallel, headphone-based 3D is experiencing ~~huge~~ an unprecedented rate of development~~growth~~. Headphones with spaced multiple drivers are emerging, accelerometers are being incorporated to facilitate head-tracking-driven audio panning, and binaural algorithms are steadily improving. The principal exponents of such playback are virtual and augmented reality. While gaming will lead this market, soon the applications will proliferate into many areas of daily life – from productivity to education, through social networking to music playback, and advertising will pervade these new media.

All of these applications require the production of audio, and this chapter explores the associated implications for the music producer. 3D offers an exciting new paradigm for music, since the conventions associated with stereo and horizontal surround are increasingly outmoded by the greater options associated with perception of height.

This chapter will consider relevant technologies and identify key production-practice, underpinned by a framework of scientific research to help define the emergent praxis of 3-D audio production. It will consider opportunities, applications and limitations,

all of which combine to introduce and help define a new approach to music production that will increasingly pervade into the future.

Introduction

For the purposes of this chapter, 3-D audio can be defined as the reproduction of sound that is perceived as coming from all around us, including behind and above, or even below. This is of course the manner in which we perceive the real world in daily life. The recreation of such perception straddles many disciplines, from recording to psychoacoustics—, to reproduction itself—, which Michael Gerzon (1973) referred to as periphony. While this chapter is centred on ‘production’ (a multifaceted term in itself), it is also necessary to contextualize this with a number of references to these various disciplines.

Clearly, when working in 3D, the most radical departure from stereo[i] or horizontal-surround production is that of source placement in the sound field, and so a perspective of this will be presented. Having said that, periphony brings other creative opportunities and hazards to the production process, and some of these will also be discussed. Further, the word ‘audio’ generally refers to music, but might be taken in a broader context to include dialogue and non-musical sounds that might accompany video.

Beyond academic research, little has been published on periphonic production at the time of writing (fall of 2017), and so this chapter also has a responsibility to introduce the topic in order to serve as a primer for the numerous texts that will doubtless be published in the future, but also to draw the reader's attention to specific aspects of the periphonic arena so that they might seek out and read further specialist texts.

There are three main paths by which 3-D immersive audio can be constructed: channel based, Ambisonic and object based~~, and these will be discussed~~. These approaches can, of course, be combined in any given playback scenario. At some point each of these requires rendering to speakers or headphones for playback. Whilst the details of each system differ and a good understanding of their workings is important, there are a number of audio-production approaches which tend to be similar across the board, although the manner of 'getting there' may require slightly different approaches.

The context of 3-D audio

Periphonic reproduction has been the subject of research for a number of decades. Over this time the degree of interest has ebbed and flowed, and various systems have been developed that facilitate so-called 'immersive audio'. It is worth first considering some of the context and various applications that employ 3-D audio.

In acousmatic music, composers have styled their music to be played ~~back~~ through multi-channel (often periphonic) speaker arrays. In 'live' performance, a musician designated as the 'diffuser' might dynamically spatialize the performance in a given array~~—~~, moving particular instruments from speaker to speaker, and many such arrays have been configured in universities and performance venues. The term 'engulfment' has been used in such circles to refer to representation of height[ii].

In the past five or six years, companies such as Dolby and Auro-3D® have developed multi-speaker periphonic playback systems for cinema. Most recent Hollywood blockbusters have embraced this approach, and consequently 3-D audio creation is now commonplace in post-production. Special effects and atmospheres can exploit the ability to be placed anywhere in a 3-D sound field around the listener, and this can

enhance synchresis[iii] and opens many possibilities for non-diegetic immersive audio. The listeners might perceive ambient sounds all around them, or experience dynamic effects such as aircraft flying overhead that can disappear into the distance behind. The associated music composition/production is also beginning to exploit such playback[iv].

The ‘soundbar’ is an increasingly popular add-on for large-screen televisions; it attempts to extend the perceived width of the sound sources beyond that of the television screen and might work both horizontally and vertically through a process termed Wave-Field Synthesis – sometimes also bouncing sound off the ceiling in order to gain the perception of height.

Binaural audio can give the impression of front-back and up-down sound-source localization on headphones. It does this by incorporating natural phenomena such as the difference in time that a sound takes to reach each ear, and the physical shape of the pinnae, which together contribute to how the brain deduces the point of origin. However, because it is dependent on the listener's physiological shape, binaural performs with varying degrees of accuracy for a given individual. Binaural localization can be achieved either through special recording techniques or via artificial ‘synthesis’. In March 2015, the Audio Engineering Society (AES) published the AES69-2015 standard that provides a framework to facilitate widespread adoption into creative content in handheld devices. This will continue to accelerate development and uptake of binaural technologies.

Virtual reality, augmented reality and mixed reality are technically different, but for convenience, will now be referred to collectively by the common acronym ‘VR’. VR represents one of the biggest technological revolutions yet to be seen, and has a

revenue stream forecast to reach $108 billion by 2021 (Digi-Capital, 2017). The attendant audio is generally binaural. Localization is greatly enhanced by the head tracking that is a feature of Head-Mounted Displays (HMD), whereby the position of a sound appears fixed relative to the natural movement of the head. Whereas VR is currently dominated by ‘early-adopter’ gaming applications (the gaming industry can be considered as residing ‘within’ VR for the purposes of this text), its impact will become more ubiquitous as it expands towards productivity, education and health. VR and its associated revenue has precipitated a step change in interest and momentum for 3-D audio.

360º video is also in ascendance. In March 2015, YouTube launched a dedicated 360º video channel, and Mobile VR (use of a smartphone with a low-budget HMD ‘holder’) is driving its popularity, and these videos typically require 3-D audio soundtracks.

Broadcasting is another industry that is increasingly experimenting with 3-D audio. Both the BBC and NHK in Japan have broadcast with a 22.2-channel system, and the BBC has increasingly created binaural content both with and without accompanying visuals, e.g. *Dr. Who* and *Radio 3*.

All of the above requires specialist equipment and thus another sector that is rapidly expanding is that of 3-D audio tool-developers, both hardware and software. Established international players such as Dolby and Pioneer are creating hardware, as are emergent start-ups such as Smyth. One interesting example is OSSIC X headphones which measure physiological characteristics of the listener's head and ears, feature spaced directional drivers to help create the sound field, and have inbuilt head-tracking. Software is coming from a similar range of developers, from the

international DTS:X protocol through to niche developers such as plugin manufacturer Blue Ripple Sound.

There is a great deal of academic research into the many areas associated with 3-D audio, spread across the humanities, science and medicine. These include music composition and performance, sonic art, human-computer interface design, audio post-production, cognition/perception, architectural space-modelling, recording techniques, software design, creative applications, assistive technologies and more. Publication is increasingly widespread, and in 2016 the AES held its first dedicated 3-D-audio conference, and The Art of Record Production held its first such event in late 2017.

Spatial placement

We must first consider some of the issues associated with 3-D audio 'placement' in reproduction. Our ability to identify a given audio source's point of origin in space – with a meaningful resolution, requires the right kind of information to be delivered to the ears, so that the brain is able to reconstruct an ostensibly correct (or at least intended) placement of that audio.

Creating a 3-D audio experience requires the presentation of usable spatial information to the ears so that this mental reconstruction can occur. The production of useful spatial information can only come via three mechanisms:

1] 'Captured' from the real world

The first is to try and capture some of the spatial cues that exist in real acoustic environments using two or more microphones, and reproduce these cues over speakers or headphones. A single monophonic microphone recording may contain a number of clues as to the environmental acoustics of the space and the sound source

being captured, but some of amount of ‘directionality’ is required if one wants to begin expressing spatial positioning from native audio recordings—, and this requires two or more microphones. Some of the inherent spatial information can then be captured and later reproduced from the timing and intensity differences between the capsules at the moment of capture.

2] Synthesized artificially

When it is not practical or desirable to capture real-world spatial information, one can model the acoustics that would be generated by a sound source in an acoustic environment synthetically. There are myriad ways of achieving this, including using spatial impulse responses, real-time room modelling and synthetic HRTF modelling, some of which will be discussed later.

3] As a function of the playback system

Panning a sound within a 3-D speaker array naturally generates spatial information at the moment of playback that is not inherent in the source audio itself. By nature of the physical relationships between the speakers, the room and the listener, 3-D positioning may be induced from power-panning alone. Good localization may thereby be achieved in a speaker array only from these factors alone. This reproduction of spatial cues will potentially be partially or completely absent from headphone reproductions of the same material.

The dynamic between these three fundamental approaches is at the center of all 3D audio production. The intended playback system (or systems) that any given 3-D audio production is expected to play back on will fundamentally influence the approach that is required. There are potentially major differences between the aesthetics and approach for producing a large-format theatrical electronic-music track

and an acoustic orchestral piece of music for a 360º video. The real challenge is where and when it is appropriate to use different capture and synthesis techniques, and where and how transforms between formats should occur. So, to elaborate~~…~~:

1] Spatial Capture

Specialized recording techniques have been designed in order to capture audio in three dimensions. One seminal approach that is currently regaining popularity is Ambisonics. This concept was developed in the 1970s by Gerzon and others, and an excellent collection of Ambisonic resources has been compiled by Richard Elen (1998).

Ambisonics can function at various resolutions that determine the accuracy of the perceived source localization, and this determines the required number of channels[v].[vi] These resolutions are categorised by 'order.'~~, and for~~ For instance, 'first order' is a four-channel system. A very important aspect is that these channels are encoded/decoded, meaning that the signals derived from (say, four) microphone capsules are not the same as those transmitted – they are encoded into a different four for transmission, and these are not the same as those passed to four speakers for reproduction – which must first be decoded (also known as rendered) from those transmitted. Alternatively, decoding could be into other formats such as 13.1 or two channels for binaural playback on headphones[vii].[viii]~~-~~ This offers a great deal of flexibility for different capture, transmission and reproduction configurations. Importantly, the transmitted signals – which are called B-format – are easily manipulated through computationally efficient multiplication by a scalar, and such operations can for instance rotate the sound field.

One reason for the resurgence in popularity of Ambisonics is that such rotation is what typically happens in a VR HMD, where the sound field appears stationary independently of the movements of the head: an accelerometer/magnetometer/gyroscope combination generates such a scalar to manipulate the B-format signals before decoding. Such dynamic panning greatly resolves the front-back confusion typical of binaural playback, and gives an overall much greater sense of realism. Such operations are typical of contemporary HMDs.

Higher-Order Ambisonics (HOA), sometimes called Scene-Based Audio (SBA) (Shivappa et al., 2016), is when additional channels are incorporated to represent further 'spherical harmonics'[ix], and these give much better localization and a larger sweet spot as the order increases. The trade-off in this approach is the increased channel count, and in the case of seventh order, 64 channels[x] are needed to convey a single track of audio, and this puts considerable strain on any Digital Audio Workstation (DAW) as the track count increases during music production. It should be remembered that beyond Ambisonics, there are other ways to capture 3-D audio too, e.g. those of Michael Williams (2004).

2] Synthesized

Where it is not possible, desirable, or practical to capture the spatial information that is present at a given recording situation, then spatialiation cues can be synthesized after the recording event.

Spatialization of monophonic audio cues can take two forms; firstly simple panning through a periphonic system that reproduces HRTF effects, and secondly, through the simulation of room acoustics that aim to localize and place the audio event realistically in 3-D space. With regard to the latter, in a real-time game engine this

acoustic simulation may be largely procedural, and can include reflections based upon room geometry, the occlusion of sounds behind physical objects, sound radiation patterns and frequency roll-off over distance – all of which are dependant on the game-software code that is deployed and the processing limitations of the system.

With linear playback materials all these effects can be generated offline by engineers using similar tools, but with more direct access and control over the materials.

The addition of reverberation to sounds has been proven to increase a sense of externalization when listening on headphones, at the expense of localization accuracy e.g. (Begault, 1992). Real-time game engines such as Unity and Unreal Engine already have such built-in audio physics, and these can be further enhanced through a number of proprietary plugins, code or middleware specifically designed for more realistic synthesis (e.g. Steam Audio, Oculus Spatializer and G'Audio etc.).

Synthetic spatialization of music elements will frequently take the form of individual monophonic sounds being placed in 3D with a specific direction and distance from the listening position. This is particularly true for more 'abstract' electronic sounds, which do not have the same restrictions of placement and acoustic realism demanded by 'acoustic music' e.g. orchestral or acoustic folk. While this approach can be enough to generate the necessary 'spread' of sounds, some judicious use of *spatial* early reflections is often helpful in placing these sounds satisfactorily.

Synthetic spatial early-reflections are timed delays that try to emulate the timing *and direction* of the first reflections to reach the listening position before the reverberant field takes over. If these are not captured with some form of multidirectional microphone array at the time of the recording, it can be useful to use synthetic versions to achieve good placement. It is possible to create such spatial reflections

with combinations of multiple delay units – if the engineer has the processing and patience necessary to set things up appropriately. Martin and King (2015) give an in-depth description of just such a system in the description of their 22.2 mix setup. There are a number of emergent bespoke systems that can also achieve this, as will be discussed. It is very important that any audio potentially destined for later synthetic spatialization must be recorded ‘clean, close and tight’ if believable results are to be achieved.

3] Playback

Sound is literally spatialized as soon as it is played back through a periphonic system by definition of the physicality of the 3-D array[xi], positioned in the array by amplitude panning.[xii] The configuration of the speakers and the acoustic properties of the room all come into play as a holistic system to make sounds appear from different places around the listeners. Similarly the configuration of the panning system also has an effect, which might follow one of a number of mathematical approaches, e.g. Ambisonic panning, Vector Base Amplitude Panning (VBAP) (Pulkki, 1997) or Distance Base Amplitude Panning DBAP (Lossius et al., 2009), together with the nature of the co-ordinates that are used to describe placement. When transforming audio between systems these issues need to be taken into consideration.

In a binaural version of the same mix, all of the effects of speaker placement, head morphology and room acoustics are typically absent or grossly simplified (although modelling is rapidly developing). Thus, if translation between headphone and speaker based mixes are required, the approach must be considered with care to either maximize ‘natural compatibility’ or allow for the various mix components to be optimally reassembled.

Objects and Channels

There are some fundamental differences between channel-based and object-based 3-D production. ‘Objects’ are aspects of audio, including both PCM-type files and associated parameters that exist discreetly, only to be reassembled at the point of playback, and this approach offers great flexibility around the modes and devices of reproduction (Pike et al., 2016). In 3-D work, a given audio file might be associated with a specific point in space and thus be linked to its spatial playback information. This approach is independent of the number of speakers or size of a given auditorium, and is employed in current 3-D cinema systems such as Dolby Atmos® as well as gaining increasing use in broadcast.

Channels (in this context not to be confused with the constituent audio streams of multichannel tracks as above) refer to the routing of an audio source to a discrete speaker output. This paradigm is familiar in stereo[xiii], and has also been traditionally used with horizontal surround systems.[xiv] Such ‘native’ recording allows for the capture and reproduction of timing differences to be conveyed between mic-capsules and speakers. These differences are crucial for convincing 3-D reproduction, and such an approach allows some of the real spatial information that is present in the recording location to be encoded and transmitted over speakers (or binaurally rendered to headphones) at a later stage.

Conversely, one cannot record natively in any meaningful way for object-based workflows. The engineer is limited to reproducing notionally monophonic recordings in some kind of monophonic-object playback/rendering system. The objects can be placed in a periphonic system using amplitude panning, but a ‘believable’ placement requires (at least) convincing early reflections since these are in major part of psychoacoustic source localization. Adequate tools for generating satisfactory

reflections (and hence externalization) in object-based systems are still developing, since until recently, the CPU cost of generating ‘scalable’ reflection/reverb systems often outstripped their commercial application, and the approach can adversely effect the balance and spectra of the mix in different rooms. De Sena et al. (2013, 2015) have made considerable progress in this regard with a computationally efficient reverb algorithm that uses a network of delay lines connected via scattering junctions, effectively emulating myriad paths that sound like propagation around a modelled room. Also, FB360 employs a system that models some early key reflections to both enhance binaural placement, and also allows the subsequent application of ‘production reverbs’ in the processing chain. Oculus and Google are currently developing near-field binaural rendering that takes account of the perceptual issues caused by head shadowing (Betbeder, 2017), and also the modelling of volumetric sounds that emanate from areas larger than a more typical point source (Stirling, 2017; Google, 2017).

One current way of approaching this reflection/object problem is to allow reflections and reverb to be generated in (channel-based) beds, whilst the objects themselves are rendered as ‘dry’ objects at playback (perhaps with monophonic early reflections/reverb ‘baked-in’). This may give a reasonable semblance of realistic positioning; however, the effectiveness of this approach is very dependent on the audio material and there may be perceptual differences at different playback locations in the sound field. Equally, different kinds of audio material have varying degrees of dependence on the reproduction of these kinds of ‘realistic’ reflections. For example, in orchestral material both the recording environment itself, and the positions of the instruments therein are crucial to the musicality of the piece, and so such music is heavily dependant on ‘authentic’ spatial reproduction. More ‘abstract’ materials (e.g.

pop or electronic) may not be as dependent on realistic acoustic placement, and so may translate equally well in an object-based environment – since the clarity and depth of each sound's position is inherently less fixed and thereby open to a less strict spatial interpretation.

There are also implications for using effects processing in channel and object-based systems. Dynamic Range Compression (DRC), as usually used in standard stereo systems starts to take on a more complex character when working spatially. Unlinked channel-based compressors create deformations of the spatial image, and there are few linked compressors that work beyond stereo or 5.1. Bryan Martin (2017) states that in 3D, DRC begins to sound 'canned' and tends towards making sound sources appear smaller rather than larger. Further, when working in pure stereo the artefacts of compression might tend to become desirable yet subtle components, but these same artefacts become increasingly apparent and unpleasant in large 3-D formats.

It is not yet feasible to have a true object-based compressor (that is, one that will operate consistently at the playback stage) and control of dynamic range – such as is possible – can only really be made on the original monophonic sounds before they enter the panning system. Having said that, the MPEG-H 3D audio-coding standard provides an enhanced concept for DRC that can adapt to different playback scenarios and listening conditions (Kuech et al., 2015). Given the increase in popularity for object-based approaches, it is likely that solutions will soon emerge that unify the object production-chain. Aside from technical challenges, keeping within acceptable processing requirements can be an issue. Systems that offer hardware rendering such as PSVR or Dolby Atmos® tend to avoid such concerns. Indeed, for theatrical applications these issues are generally less problematic, since cinema sound tends to not rely on the *character* that compression effects bring to material, but for creative

music production, the compromised nature of this familiar toolset tends to have greater implications.

Indeed, these issues apply to any traditionally insert-type effects that might affect objects spectromorphologically. It is also worth noting that because Ambisonic audio is encoded throughout the production chain between the point of capture and reproduction, it is not generally possible to apply any kind of effects processing to it either, since such ‘distortion’ will conflict with the delicate spherical harmonics, especially at higher orders. There are a few ambisonic plug-ins that offer EQ, delay and reverb, but these tend to be limited to lower-order operation.

As a side note, it is worth noting that object-based production workflows are nothing new per se. At all times during a DAW-based mix, one is effectively working with objects. The fundamental difference is that during ‘normal’ mixing the rendering to speakers happens at the same moment as the mix proceeds, as opposed an object-based mix that postpones that rendering to the moment of playback at the consumer end. An obvious implication in addition to the above, is that the normal kinds of ‘buss’ processing which engineers are used to are no longer available when the mix is fragmented into its constituent objects, since the summing of the playback/production chain has been deferred to the moment of reproduction. This is why we currently see hybrid systems of objects and channels, and ‘pure’ object-only systems are still in evolution. Modern game engines also follow this logic – and in fact in many ways have been leading such workflow architecture for some time – blending stereo and 5.1 tracks with real object-based sounds to achieve pleasing results.

Capture considerations

When assessing the value of the channel/object parts of a given system, one must look at the material conditions of the recording in question. If one wants to capture something of the 'acoustic' and the spatial relationships between elements as they were in the room at the time of recording (and have real coherence between speakers at playback) then some form of multi-microphone/channel-based system currently tends to be preferable to the object approach (at present). This state of affairs is likely to be the case only so long as synthetic acoustic modelling tools lag behind the 'real thing'. As soon as the audio equivalent of photo-realistic visual effects is achievable, the flexibility advantages of an object approach may make it the dominant form.

It is worth considering the difference between working with spaced microphones and an Ambisonic microphone (e.g. a Calrec Soundfield). Of the former, Williams' psychoacoustic research (2004) looks at the way in which multi-microphone systems create 'coverage' between any given pair of capsules. Williams notes that the Stereo Recording Angle (SRA), coverage, and angular distortion[xv] are all dependant on a combination of the microphones' polar patterns, their relative angles, and their distance from each other, and this points to potential issues with Ambisonic microphones that are frequently overlooked ~~(Williams, 2017)~~. An Ambisonic microphone utilizes (theoretically) coincident capsules, and this means that it can only record intensity *difference*. Crucial timing information from the recording location is lost (and cannot be recovered). Further, the capsule types and angles imply large sections of SRA overlap in the horizontal plane once summed for playback, and this gives rise to angular distortion at regular intervals around the full circular field of capture (which becomes increasingly apparent outside of a very small sweet spot). The lack of timing information, coupled with the overlap of polar patterns can lead to

an unsatisfactory image that lacks the spaciousness and accuracy of a spaced-array recording (Williams, 1991).

Some of the issues of first-order ambisonic microphones have been improved with the recent introduction of higher-order ambisonic microphones. These greatly improve some of the angular distortion issues associated with first-order variants, but they can come at the cost of the spectral-colouration issues such as the low frequency roll-off associated with more directional cardioids, or potentially high-frequency phase artefacts from their large number of capsules. As is so often the case, there is a trade-off to be made between spectral and spatial quality on the one hand, and practicality and cost on the other, and the engineer must evaluate this for each project. New tools are emerging to support this, such as Hyunkook Lee's (2017) Microphone Array Recording and Reproduction Simulator (MARRS) app, which predicts "the perceived positions of multiple sound sources for a given microphone configuration" and can also "automatically configure suitable microphone arrays for the user's desired spatial scene in reproduction".

In a VR workflow there is often an assumption that Ambisonic recording is 'native' recording. Whilst this may be true enough for some kinds of single-perspective 360º video recordings, it can be less so for 'true' VR that utilizes 6DOF[xvi] - where the listener's locative movement within a virtual space and its associated sound field requires a shift in sonic perspective.[xvii] Spaced arrays and multi-microphone setups can be reproduced in VR if care is taken over their reproduction in the audio system of a game engine. Issues with phase tend to be dependent both on the material being reproduced, the placement of the 'array', and the particular qualities of the binaural decoder. Recordings with low levels of correlation in each channel tend to work best. Game engines also allow for other more sophisticated ways of placing and

spatializing monophonic sounds, and Ambisonics might be only one aspect of the greater sound mix, perhaps layered with binaurally synthesized elements and head-locked audio[xviii]. Further, coverage from multiple microphones can be blended into ‘scenes’ using synthetic acoustic placement (see below). Summing synthetically positioned spot-microphones and ‘spatial’ Ambisonic recordings can achieve very satisfactory results and becomes necessary to satisfy the requirements of balancing dry/close recordings with more distant or reverberant elements. This approach has been corroborated by Riaz et al. (2017) in a scientific setting, although the full results of that test are not published at the time of writing. Phase is likely to play an important part in the success of such ‘scene’ combinations and so the engineer might evaluate this and adjust it for optimal timbre and placement.

At this time, a common problem with HOA is that of channel count, since the host DAW perhaps needs to accommodate say 36 (for fifth-order Ambisonic) channels for each ‘monophonic’ track of audio, and this quickly becomes demanding in terms of CPU load and streaming bandwidth. Mixed-order Ambisonics can offer a solution whereby different orders are combined, perhaps with first-order Ambisonics for sounds like beds[xix] without need of more precise localization, and HOA for key sounds in need of accurate placement. This approach might also be implemented by providing decreased resolution in the vertical plane to which the ear is less sensitive; the ears sit on a horizontal plane and are therefore more attuned to it (Travis, 2009). However, in either case, care is needed with the type of decoder used:

“A 'projection' decoder will have matrix coefficients that do not change for the first four spherical-harmonic components for 1st, 2nd and 3rd order. So simple summation in the equivalent of the ‘mix-buss’ will work here, regardless of the layout. In contrast, a pseudo-inverse decoder won't behave this way and will have different

matrix coefficients for the first four spherical-harmonic channels for each order. Thus the layout must be as symmetrical/even as possible, for this to give results similar to projection." (Kearney, 2017)

When working in linear environments like cinema (as opposed to real-time, for instance in VR or gaming, where spatialization requirements might change dynamically) there is scope to render high-quality acoustic spatial elements. For instance, it is possible to ameliorate the above CPU and bandwidth issues by 'baking-in' higher-resolution reverbs into higher-order Ambisonic beds. Another approach taken by the PSVR system uses an object-based approach, sending each voice with its associated azimuth and elevation information, via middleware to bespoke audio libraries in the PSVR hardware for rendering along with mixed-order components. This gives excellent localization for the objects – which can also be moving, if tied to animated 'emitters' associated with visual elements of a game or application – but loses the ability to perform more typical channel-based production processing en route.

Mixing for Binaural

The pinnae, head and torso modify sounds before they enter the ear canal (filtering via reflection effects) in unique ways that are individual to the listener and dependant on the morphology of *that* person. Whilst ITD (Inter-aural Time Difference) and ILD (Inter-aural Level Differences) are important for determining the position of sounds in the horizontal plane, spectral content, filtered largely by the pinnae, is mostly responsible for the perception of elevation, or sounds in the median plane[xx] (Wightman and Kistler, 1997), although this is slightly augmented by head and torso effects.[xxi] Combining the above into filters whose response is given by such a Head-Related Transfer Function[xxii] (HRTF)[xxiii] allows a sound to be processed[xxiv] by that

filter in order that the brain perceives it as coming from a particular point in space.[xxv] Such HRTFs need to be provided for every point in space that might be emulated, and the appropriate HRTF filtration applied to a source sound relative to a given localization placement. This is the mechanism of binaural reproduction in headphones and forms the basis for the binaural 'synthesis' previously mentioned. Although HRTFs are person-specific, there have been a number of one-size-fits-all ones developed, e.g. Furse's (2015) 'Amber' which exploits statistical averages to gain the best (compromised) optimization for the largest number of people. A good introductory treatment of binaural theory is given by Rumsey (2001).

The implications of the above for recording and mixing in 3D are important. The first is that binaural[xxvi] 3-D mixing on (for) headphones, using a non-personalised off-the-shelf HRTF will tend to lead to mixes which may or may not work as intended for others depending on the relative differences between the morphologies of: 1) the mixer 2) the HRTF model that was utilized, and 3) the end listener. Whilst some tools allow the engineer to select a preferred HRTF, although this will undoubtedly give a truer and more responsive experience to that individual, it is less likely to successfully translate to the largest proportion of end listeners. One option is to first mix to a favoured HRTF, then switch to one such as Amber and then optimize the mix for that prior to final distribution. Clearly, such an approach is an idiosyncratic choice.

Another strategy is to mix as much as possible over a 3-D speaker array and use a conversion transform at the end of the mix process to complete the 3-D processing to a headphone format. This reduces the error distance between the mixer's morphology and the end user, since the engineer will not attempt to correct the HRTF model to suit his or her own peculiarities. If this approach is adopted then care must be taken to ensure that the conversion tools are high quality, since there can be large differences

in the quality of encoder/decoders that convert between speakers and headphones, and vice versa. This approach can represent something of a leap of faith for the engineer though, who simply has to accept the algorithmically created binaural mix. There is some hope that more personalized HRTF models will ameliorate some of these margins of error in the medium term, but it remains an issue at the point of writing.

There are numerous sets of Ambisonic software tools available at the time of writing, many of which are freeware. Several also feature binaural decoders, and can therefore monitor and render for headphones. Many of these tools tend to be first order, although the excellent ambiX suite (Kronlachner, 2014) goes up to 7th order.

A typical DAW workflow would be to insert an Ambisonic panner (an encoder) onto a monaural audio track, and the panner would be able to send its multichannel output to a buss, with the track's main output muted. The panner will convert the track to (say) 4-channel for first order – which can now accommodate 3-D spatial information. A similar setup should be implemented in other tracks. The four-channel busses should be routed to another track of the same channel width, and a suitable binaural decoder should be inserted on this track. This will have four inputs and two outputs, the latter representing the headphone feed. The inserted panners can then be used to spatialize the various tracks, thus localizing monophonic sound. This basic setup can be extended by having a parallel destination track with an Ambisonic reverb inserted on it, then fed into the main destination track prior to the decoder. To work at higher orders, panners and decoders designed for this can replace those above, and the channel count increased accordingly for all tracks and busses.

People will perceive the output of such a system in different ways to different degrees of the intended effect, a function of the HRTF issue, but also variations in headphones including their physical structure[xxvii] and more. It is generally possible to create

definite localization to the rear, although sometimes this is just perceived as ultra-wide stereo. Elevation tends to be less successful. However, interesting mixes can be created with good separation and width. Of course, the engineer cannot know exactly how others will perceive it.

As mentioned, much of the current interest in in binaural Ambisonic audio is coming from VR HMDs, and crucially these feature head tracking. As mentioned, an Ambisonic sound field can be rotated through simple multiplication by a scalar, and if head-tracking data is converted to such a scalar, then the sound field can be rotated in counter motion to the head to give the impression of sounds that are fixed in space and independent of the listener's head movement, as is of course the case with real-life. Such head tracking is very effective at removing front-back confusions, and further greatly enhances perception of elevation, even if that is only through the listener being able to ‘look up’ at the audio and bring it into frontal auditory focus to affirm that it is up there. To monitor this in a DAW, a ‘rotator’ plugin can be inserted before the decoder, and head-tracking data routed to the rotator’s parameters[xxviii].[xxix] To replicate the effect for the end user, an HMD needs to be supplied with B-format Ambisonic[xxx] multi-channel audio so that the rotation can be rendered ‘live’.[xxxi] At the time of writing, various head trackers are available that can be attached to normal headphones bypassing the need for an HMD, although they typically have various compatibility issues, for instance being tied to a given manufacturer’s software. Hedrot (Baskind, n.d.) is a cheap DIY head tracker that is widely compatible via OSC data.

The nature of HRTF-based spatial modelling means that filtering (particularly of upper frequencies) is readily discernable in many listening directions. This becomes much more apparent with dynamic head tracking – manifesting itself as something of

a sweep, and so care must be exercised regarding what elements of a mix are panned where. If for instance a drum-kit stem is placed at the front, as the head turns there will be noticeable attenuation of the cymbals as the head is turned, and this might be less preferable than the desired effect of spatialization. Head locking[xxxii] of such parts provides a solution. Sounds with more arbitrary timbre are more tolerant of such panning since there is a less polarized 'right' sound, and so these might be better placed around the sound field. Good effect can be gained by fixing a key sound such as bass to the front, since the anchored low frequency energy gives a strong impression of rotation regardless of what else is panned, but care must be exercised since rotating low frequencies can also induce nausea, and so judicious filtering – possibly splitting a given musical part into separately locatable tracks – might be necessary. Interesting effects can be readily created, for instance with high-pass filtered reverb only apparent when looking up, which can give the impression of a high ceiling. Something that should be considered is that a unique musical journey can be created dependent on head position. The music might literally be presented differently dependent on the user's aural 'field of view', analogous to visuals on a train journey where one might look in different directions each time the journey is taken. It is quite possible that head tracking will become a standard feature of many headphones, perhaps especially as Augmented Reality proliferates.

Some speaker considerations

It is commonly accepted that vertical perception is guided by spectral cues, whereby the higher the frequency, the higher the perceived height relative to a single source loudspeaker, and in fact for a center speaker (on the median plane), frequencies below 1 kHz tend to be received as coming from lower than the speaker. This is supported by several scientific studies and is called the pitch-height effect (Cabrera and Tilley,

2003). Lee (2015) developed an upmixing[xxxiii] technique called Perceptual Band Allocation (PBA) that could enhance the 3-D spatial impression by band-splitting horizontal-surround-recorded ambience and routing the bands to the higher and lower loudspeaker layers of an array (the upper frequencies to the upper layer).[xxxiv] The PBA system outperformed an eight-channel 3-D recording in listening tests, and a center frequency of 1 kHz was proposed for this to work[xxxv].[xxxvi] Lee (Lee, 2016a, 2017a) also examined the degree of Vertical Image Spread (VIS) – the overall perceived height of the sound field, this time with phantom images. This was also done via PBA, and he experimented with the control of the VIS on a per-band basis by mapping multiple frequency bands to speaker layers. Lee found that the vertical locations of most bands tended to be higher than the height of the physical speaker layer and that it was generally possible control the VIS via different PBA schemes.

There is another psychoacoustic phenomenon related to the pitch-height effect; the 'phantom-image elevation effect' (De Boer, 1947), that also influences our perception of elevation. This is the observation that for perfectly coherent stereoscopic phantom-center images on speakers (e.g. a vocal, panned centrally between a pair, but not routed directly to any center speaker), different frequencies perceptually 'map' themselves to different elevations in the median plane. It is a common approach in channel-based mixing with a speaker array to employ a conventional stereo source routed to a pair of speakers, which of course might typically form a phantom image in the center. In this context, Lee (2016b, 2017b) again explored the relationship between frequency and psychoacoustic perception of height using a variety of loudspeaker pair base-angles, including the conventional (stereo-like) 60°, and 180° pairs that were orthogonal to the median plane – directly to the either side of the head. For the 60° pair, broadband frequencies appeared to be elevated, and for the 180° set-

up, both octave bands of 500 Hz and 8 kHz, and sounds with transient characteristics were the most prominently elevated, and the 1 kHz region was often perceived to be behind the listener for all base angles. Further, as the base angle increases from 0° to 240°, the perceived phantom image is increasingly elevated. Overall, elevation was negligible (along with general directionality) for frequencies below 100 Hz. Although the higher frequency elevations aligns with Blauert's (1969) 'directional bands theory', Lee hypothesizes that the sub-3 kHz effect is due to acoustic crosstalk from a pair of speakers being interpreted by the brain as reflections from the shoulders.

The application of such theory to mixing offers an enhanced understanding and extends the possibilities for spatial placement. When mixing to 'single' speakers, EQ-ing to roll off above 1 kHz can extend the sound field to below that speaker, and in general, higher frequencies will be perceived as coming from progressively higher. Although the pitch-height theory pertains to the center speaker, this effect might momentarily extend to any speaker in the array since as the listener rotates their head to face it, it comes into the effective median plane. Sounds panned to the height layer of a speaker array that contain a certain amount of higher frequencies may be perceived as emanating from even higher, offering an extension of the sound field, and high-pass filtering might offer a further exaggeration of this effect.

Phantom centers present interesting opportunities. When working with 60° stereo pairs of speakers that create a coherent phantom center, most frequencies will elevate, and these might be balanced against any signal actually routed to the center speaker to create a vertical panorama on the median plane. Mid-range content (possibly band-pass filtered, since the 250 Hz octave elevated less) panned equally to a 180° pair will again elevate the phantom image, and this could be implemented with either a main speaker layer or indeed the height layer to gain greater sonic elevation; the same

applies for 8 kHz. Increasing elevation might be generated by automating the routing to progressively wider pairs of speakers, but this should be done incrementally rather than progressively to avoid comb filtering and transient smearing in the process. The 1 kHz octave band has a ‘center of gravity’ behind the listener, and thus panning and EQ can be arranged to exploit this. Conversely, it is useful to remember not to ‘fight’ this phenomenon.

Also, where bespoke 3-D reverbs are not available, one can follow the PBA ideas and achieve good results by band splitting a reverb at, say 1 kHz, and placing the hi-passed components in the height layer. This expands the vertical perception of the space, and has less of the problems of using two reverbs to achieve the same result.

Thus, these theories present a basic premise for working with height in certain circumstances, although naturally any such spectral presentation is likely to be subservient to the greater tonality of the mix. It should be remembered that much of this applies principally to phantom-center images, and other components of a mix might easily compensate – or fight – spectrally and hence spatially. There is a copious body of literature on 3-D capture for the interested reader to explore; the archive of the Audio Engineering Society is a good place to start.

Mixing: general approaches

As one moves towards higher spatial resolution in 3D, the intrinsic quality of the original component recordings/parts of the piece become increasingly laid bare, allowing a greater focus on the individual elements of a mix. These now have more space to be heard clearly, just as might be desired by three-point panning in stereo. That aside, in stereo, when all sounds come from just a pair of speakers there is a lot of spectral and transient masking & blending that can greatly obfuscate many of the

elements within that mix. This can of course be a very desirable aspect of balancing the competing elements of a complex music mix, and may in fact be the reason that 'mixing' might be required in the first place. As the speaker count (or binaural headphone resolution) increases, so do the opportunities to more precisely discern individual elements (assuming that both the total number of audio elements in the mix remains constant and also exploits the extra 'space'). As this happens, the otherwise 'hidden' qualities of the underlying elements are increasingly exposed – for better or for worse. This is doubly so when utilizing monophonic sounds in an object-based system, where background noise and recorded reflections can compromise the ability to repurpose the sound in 3-D space. Bryan Martin (2017) attests to all this and also emphasizes that edits must be much tighter than are often accepted in dense stereo work.

For these reasons, 3-D mixing invariably demands a number of new things of the mixer:

1] There needs to be a greater amount of sonic material to 'fill' the new-found space, and this can be approached in two ways; from a compositional/arrangement perspective, or from a pure *mix* perspective. Clearly, compositionally there is physically more *space* for the arrangement of elements in a 3-D space than there is in a stereo mix. Adding new layers of instrumentation (which need not represent harmonic extensions of the extant musical parts) is one approach to fill this space, although this approach overtly changes the character and arrangement of the original music. The same approach works more readily in film where ambience can more easily be added to and augmented into height/surround layers based on the content of the original stereo ambience. With music, this approach can lead to complexity and adornment that may be undesirable, and it is an approach that only tends to works

with abstract (non-acoustic) performances. Of course, many composers will relish the opportunity to embrace such a medium and will tailor their arrangements in order to exploit the new-found possibilities. As 3-D delivery proliferates, such approaches will likely become accepted and commonplace.

Perhaps the more readily adopted approach is to use mix techniques to expand the sonic space of the original, in ways that do not compositionally change the piece in any radical manner. There are a number of approaches to achieving this kind of 'spread' of the component elements. The first is to simply pan elements to their new positions in the 3-D sphere. The downside of this is that in their own right, each direction's audio can appear somewhat spartan in content, as the original busy stereo field is decomposed into individual elements – separated and exposed over the larger playback area.

The second is to add width and size to those elements. If the placement of a sound renders it too small in the overall soundstage, when for instance it is emanating only from a single speaker, then the mix engineer might wish to spread it over more than one unit. This can have a number of effects. If it is (say) evenly panned between two speakers that subtend around 60º to the listener[xxxvii] then it will generate a new phantom center that will not necessarily sound any bigger than the original – just displaced, and biased panning will simply move the phantom's origin. Further, if transient material is present, then smearing can occur at certain points in the listening area, and this of course will be a function of the relative distance to each of the two speaker units.

If sounds are shared over two or more speaker channels it is always a good idea to ensure there is some kind of decorrelation, so that the sound is not exactly the same in all speakers – this at least mitigates the phantom-center problem, although not the

transient smearing. Sound sources in the real world are essentially decorrelated through reflections and asymmetries of the listening environment. Decorrelation within a mix environment can be achieved through slightly offsetting EQ settings, delay times, or pitch. Such EQ may not be desirable tonally, and strategic decisions might need to be taken to balance such tensions. Decorrelation can also be achieved by use of reverb systems and Impulse Response convolutions. Some commercial spatialization systems feature a parameter called ‘spread’ or some such that can automatically change the size of a source in the sound field although their effect is not always as intuitively expected.

2] Time-based effects (at least delays and echoes) can be most effective if adapted to spatially modulate over the front-back and median planes. There are also some excellent ‘3D’ reverbs that have come onto the market that can achieve great results, but stereo or mono algorithms may be used in multiple channels, so long as the settings are adjusted in each to ensure a satisfactory level of decorrelation, and of course interesting effects can be generated by separating the ‘wet from the dry’ spatially, to varying degrees.

However, delays and phase shifts can also be used to place sounds in ‘the sphere’ according to the precedence effect; there are situations where it is better not to rely solely on amplitude panning alone. The ‘ambience’ around a source might need to change accordingly – particularly for moving sources, and to achieve this without overtly muddying the mix is a question of balance, taste and technique. Smaller ‘room’ settings with short early-reflection times and shorter reverb tails will often still work well if too much reverb is problematic for the greater mix.

Martin (2015) makes a detailed examination of the use of synthetically generated early reflections in a 22.2 channel-based system, and gives clear descriptions of the

utilized timings. There is some disagreement over how well the precedence effect works in the vertical, but it is a viable option for placing sounds and the individual mixer will have to determine the impact of such use in their own mix. If translation to binaural might be later required, then the effectiveness of such panning mechanisms and indeed the reverbs must also be considered in that context. Reverbs in particular can cause problems when ‘collapsed’ from speakers into a binaural render.

3] DRC in HOA has spatial and spectral artefacts that can be most undesirable, and such a process falls into the category already mentioned that will corrupt spherical harmonics leading to a degradation of the spatialization. Also, if working with objects, it should be understood that they are by their nature hard to compress in any natural sounding fashion – just as sounds in the real world are not naturally compressed.

DRC and EQ in both 3-D speaker based systems and 3-D binaural headphone mixes must be approached with more caution than in a stereo or even surround system. As the 3-D resolution increases, the artifice of heavy compression and EQ become increasingly revealed. Martin (2017) warns mixers who wish to embark on 3-D channel based mixing to: “forget compression – it’s the opposite of what you want”. He also described the counter-intuitive experience of using stereophonic compression techniques in a 3-D-mix environment:

“Compression fails – because it makes it even smaller. You have this huge panorama to present, and compression just makes it even smaller – so it actually just fights you. People think compression is going to make it more ‘there’ but it makes it smaller, when you want it to be bigger. Compression is not your friend.”

Use of DRC as a ‘glue to hold it all together’ becomes increasingly problematic in 3D. Coupled with the increasing bias towards creating a realistic (or at least

believable) sound field around the listener that is engendered when working in higher resolution 3D, traditional ‘stereo buss’ techniques such as buss compression feel increasingly out of place – not to mention technically unachievable at present.

Richard Furse’s Blue Ripple toolset (2017) has an interchange plugin which can transform an Ambisonic mix to a 20 channel A-format that allows the use of standard stereophonic-type effects processing, followed by a re-encode to HOA. Although it is preferable to remain in the Ambisonic realm when mixing ‘Ambisonically’, Furse’s approach allows for signal processing that is not currently possible with the currently available Ambisonic toolset. Perhaps the future of mixing in high-resolution 3D will move towards increasing reliance on the highest quality reproduction of the musical elements themselves, within a dynamic and intelligent physical environment with metadata-linked ends of the processing pipeline.

Conclusion

Immersive audio indicates by its very name that some kind of presence in the performance is anticipated for the listener. This also implies that the overall effect of the sound elements should have a clarity and some kind of all-encompassing physical expression free from the limitations of a 60° wide front-facing stereo track. It is not always obvious quite how detached a lot of stereo music-reproduction is from sounds in the physical world, perhaps most greatly exemplified by classical music, and this leads to issues when extending and translating stereo techniques to 3D, which invariably lean towards bringing sounds into a more open, physically encompassing expression in three dimensions.

There are a number of approaches to capturing and creating 3-D audio, and each has its merits and limitations. While stereo & horizontal-surround music production is an extremely well established art-form with a number of conventions and aspects of

accepted good practice, many of these are subverted by the increased complexity, both perceptual and technical that nascent 3-D production entails.

As was the case with 5.1, it is unlikely that multi-speaker setups will routinely make it into the domestic living room with appropriate configuration, but as wave-field synthesis, transaural binaural reproduction and distributed-mode loudspeakers evolve, so will home 3D. However, the VR industry is expanding rapidly and bringing head-tracked audio into the mainstream expectation, and regular audio-only head-tracked headphones will soon follow in their droves. This will turn relatively banal binaural listening into something much more dynamic, and while the limitations of such systems will preclude their usage in lots of music, even small aspects of dynamic binaural reproduction will make it into a large proportion of mainstream popular music mixes. Although binaural is at the mercy of HRTF-matching at present, future solutions will overcome this and emancipate headphone listening further still.

3-D music production is evolving and will deserve increasing attention in order to develop the art form, and it is not possible to comprehensively cover the topic in a single chapter. Doubtless, much will be written in due course, and accordingly we can look forward to developing increased understanding of the praxis. The creative possibilities are exciting – presenting one of the biggest single opportunities to forward music-production in a long time. These possibilities await only innovative individuals with suitable tools to build a new sound world that will represent a step change in the creation and consumption of music.

Acknowledgements

The authors would like to thank the experts who graciously contributed interviews for this text: ~~Richard Furse,~~ Hyunkook Lee and~~,~~ Bryan Martin ~~and Michael Williams~~.

Special thanks go to Hyunkook Lee for the guidance around his own work.

References

Barrett, N. (2002) 'Spatio-musical composition strategies', *Organised Sound*, 7(3), pp. 313–323. doi: 10.1017/S1355771802003114.

Baskind A (n.d.) Hedrot. Available from: https://abaskind.github.io/hedrot/ (accessed 11 July 2017).

Begault DR (1992) Perceptual Effects of Synthetic Reverberation on Three-Dimensional Audio Systems. *J. Audio Eng. Soc* 40(11): 895–904. Available from: http://www.aes.org/e-lib/browse.cfm?elib=7027.

Betbeder L (2017) Near-field 3D Audio Explained. Available from: https://developer.oculus.com/blog/near-field-3d-audio-explained (accessed 11 October 2017).

Blauert J (1969) Sound localization in the median plane. *Acta Acustica united with Acustica* 22(4): 205–213.

Cabrera D and Tilley S (2003) Vertical Localization and Image Size Effects in Loudspeaker Reproduction. In: *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. Available from: http://www.aes.org/e-lib/browse.cfm?elib=12269.

De Boer K (1947) A remarkable phenomenon with stereophonic sound reproduction. *Philips Tech. Rev* 9(8).

De Sena E, Hacıhabiboğlu H and Cvetković Z (2013) Analysis and Design of Multichannel Systems for Perceptual Sound Field Reconstruction. *IEEE Transactions on Audio, Speech, and Language Processing* 21(8): 1653–1665.

De Sena E, Hacıhabiboğlu H, Cvetković Z, et al. (2015) Efficient synthesis of room acoustics via scattering delay networks. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 23(9): 1478–1492. Available from: http://dl.acm.org/citation.cfm?id=2824192.2824199 (accessed 3 October 2017).

Digi-Capital (2017) After mixed year, mobile AR to drive $108 billion VR/AR market by 2021. Available from: http://www.digi-capital.com/news/2017/01/after-mixed-year-mobile-ar-to-drive-108-billion-vrar-market-by-2021/ (accessed 16 July 2017).

Elen R (1998) The ambisonic motherlode. Available from: http://decoy.iki.fi/dsound/ambisonic/motherlode/index (accessed 30 August 2017).

Furse R (2015) Amber HRTF. *Blue Ripple Sound*. Available from: http://www.blueripplesound.com/hrtf-amber (accessed 7 November 2017).

Furse R (2017) 3.1 O3A B->A20 and O3A A20->B Converters. *O3AManipulators_UserGuide*, User Guide. Available from: http://www.blueripplesound.com/sites/default/files/O3AManipulators_UserGuide_v2.1.4.pdf.

Gerzon MA (1973) Periphony: With-Height Sound Reproduction. *Journal of the Audio Engineering Society* 21(1): 2–10. Available from: http://www.aes.org/e-lib/online/browse.cfm?elib=2012 (accessed 29 March 2017).

Gerzon MA (1985) Ambisonics in Multichannel Broadcasting and Video. *Journal of the Audio Engineering Society* 33(11): 859–871. Available from: http://www.aes.org/e-lib/browse.cfm?elib=4419 (accessed 29 March 2017).

Google (2017) Resonance Audio. *Google Developers*. Available from: https://developers.google.com/resonance-audio/ (accessed 7 November 2017).

Jot J-M, Larcher V and Pernaux J-M (1999) A Comparative Study of 3-D Audio Encoding and Rendering Techniques. In: Audio Engineering Society. Available from: http://www.aes.org/e-lib/browse.cfm?elib=8029 (accessed 29 March 2017).

Kearney G (2017) Ambisonic Question: E-mail.

Kronlachner M (2014) ambiX v0.2.7 – Ambisonic plug-in suite | matthiaskronlachner.com. Available from: http://www.matthiaskronlachner.com/?p=2015 (accessed 28 March 2017).

Kuech F, Kratschmer M, Neugebauer B, et al. (2015) Dynamic Range and Loudness Control in MPEG-H 3D Audio. In: *Audio Engineering Society Convention 139*. Available from: http://www.aes.org/e-lib/browse.cfm?elib=18021.

Lee H (2015) 2D-to-3D Ambience Upmixing based on Perceptual Band Allocation. *J. Audio Eng. Soc* 63(10): 811–821. Available from: http://www.aes.org/e-lib/browse.cfm?elib=18044.

Lee H (2016a) Perceptual Band Allocation (PBA) for the Rendering of Vertical Image Spread with a Vertical 2D Loudspeaker Array. *J. Audio Eng. Soc* 64(12): 1003–1013. Available from: http://www.aes.org/e-lib/browse.cfm?elib=18534.

Lee H (2016b) Phantom Image Elevation Explained. In: *Audio Engineering Society Convention 141*. Available from: http://www.aes.org/e-lib/browse.cfm?elib=18468.

Lee H (2017a) Interview with Hyunkook Lee. Interviewed by Gareth Llewelyn for Producing 3D Audio.

Lee H (2017b) Sound Source and Loudspeaker Base Angle Dependency of Phantom Image Elevation Effect. *J. Audio Eng. Soc* 65(9): 733–748. Available from: http://www.aes.org/e-lib/browse.cfm?elib=19203.

Lee H, Johnson D and Mironovs M (2017) An Interactive and Intelligent Tool for Microphone Array Design. In: *Audio Engineering Society Convention 143*. Available from: http://www.aes.org/e-lib/browse.cfm?elib=19338.

Lossius T, Balthazar P and de la Hogue T (2009) DBAP – Distance Based Amplitude Panning. In: *International Computer Music Conference (ICMC)*, Montreal.

Lynch, H. and Sazdov, R. (2011) 'An Investigation Into Compositional Techniques Utilized For The Three-Dimensional Spatialization Of Electroacoustic Music', in *Proceedings of the Electroacoustic Music Studies Conference, Sforzando! Electroacoustic Music Studies Conference*, New York, USA. Available at: http://www.ems-network.org/spip.php?article328 (Accessed: 10 February 2018).

Martin B (2017) Interview with Bryan Martin. Interviewed by Gareth Llewelyn for Producing 3D Audio.

Martin B and King R (2015) Three Dimensional Spatial Techniques in 22.2 Multichannel Surround Sound for Popular Music Mixing. In: *Audio Engineering Society Convention 139*, Audio Engineering Society. Available from: http://www.aes.org/e-lib/online/browse.cfm?elib=17988 (accessed 1 July 2016).

Martin B, King R, Leonard B, et al. (2015) Immersive Content in Three Dimensional Recording Techniques for Single Instruments in Popular Music. In: *Audio Engineering Society Convention 138*, Audio Engineering Society. Available from: http://www.aes.org/e-lib/online/browse.cfm?elib=17675 (accessed 1 July 2016).

Pike C, Taylor R, Parnell T, et al. (2016) Object-Based 3D Audio Production for Virtual Reality Using the Audio Definition Model. In: Audio Engineering Society. Available from: http://www.aes.org/e-lib/browse.cfm?elib=18498 (accessed 21 March 2017).

Pulkki V (1997) Virtual Sound Source Positioning Using Vector Base Amplitude Panning. *J. Audio Eng. Soc* 45(6): 456–466. Available from: http://www.aes.org/e-lib/browse.cfm?elib=7853.

Riaz H, Stiles M, Armstrong C, et al. (2017) Multichannel Microphone Array Recording for Popular Music Production in Virtual Reality. In: *Audio Engineering Society Convention 143*. Available from: http://www.aes.org/e-lib/browse.cfm?elib=19333.

Rumsey F (2001) *Spatial Audio*. Oxford ; Boston: Focal Press.

Shivappa S, Morrell M, Sen D, et al. (2016) Efficient, Compelling, and Immersive VR Audio Experience Using Scene Based Audio/Higher Order Ambisonics. In: Audio Engineering Society. Available from: http://www.aes.org/e-lib/browse.cfm?elib=18493 (accessed 22 May 2017).

Stirling P (2017) Volumetric Sounds. Available from: https://developer.oculus.com/blog/volumetric-sounds (accessed 11 October 2017).

Travis C (2009) A New Mixed-Order Scheme for Ambisonic Signals" — Ambisonics. File. Available from: http://ambisonics.iem.at/symposium2009/proceedings/ambisym09-travis-newmixedorder.pdf/@@download/file/AmbiSym09_Travis_NewMixedOrder.pdf (accessed 10 October 2017).

Wightman FL and Kistler DJ (1997) Monaural sound localization revisited. *The Journal of the Acoustical Society of America* 101(2): 1050–1063. Available from: https://doi.org/10.1121/1.418029.

Williams M (1991) Microphone Arrays for Natural Multiphony. In: *Audio Engineering Society Convention 91*. Available from: http://www.aes.org/e-lib/browse.cfm?elib=5559.

Williams M (2004) *Microphone Arrays for Stereo and Multichannel Sound Recordings*. Editrice Il Rostro.

~~Williams M (2017) Interview with Michael Williams. Interviewed by Gareth Llewelyn for Producing 3D Audio.~~

i Perhaps ironically in the context of this chapter, ‘stereo’ comes from the ancient Greek ‘stereos’, which means solid – with reference to three dimensionality, albeit with regard to the formation of words.

ii A body of research has formed around this, for example (Barrett, 2002) and (Lynch and Sazdov, 2011).

iii Synchresis is the psychological linking between what we might see and hear when such events occur simultaneously.

iv Media composers such as Joel Douek and Michael Price are notable for embracing 3-D approaches.

v In fact, conventional stereophony is a subsystem of Ambisonics (Gerzon, 1985)).

vi In fact, conventional stereophony is a subsystem of Ambisonics (Gerzon, 1985)).

vii Jot, Larcher and Pernaux (1999) provide a useful text on encoding/rendering.

viii Jot, Larcher and Pernaux (1999) provide a useful text on encoding/rendering.

ix Spherical polar-coordinate solutions to the acoustic wave equation.

x If the channel order = N (which defines the angular resolution of the spherical harmonics), the number of audio channels required for that order is given by $(N+1)^2$.

xi Although this also applies to any multi-speaker setup, be that stereo or horizontal planar surround.

xii Although this also applies to any multi-speaker setup, be that stereo or horizontal planar surround.

xiii e.g. in stereo: record an acoustic guitar with a single microphone and pan it to play back from a single speaker, or indeed centrally, when it is mapped to both speakers with equal amplitude etc..

xiv e.g. in stereo: record an acoustic guitar with a single microphone and pan it to play back from a single speaker, or indeed centrally, when it is mapped to both speakers with equal amplitude etc..

xv Williams used this term to describe how changes in the relative angles of microphone capsules might shift the perceived position of the instruments on the horizontal plane, once reproduced.

xvi 6DOF: Six Degrees of Freedom, which refers to movement in a 3-D-space: front/back, left/right, up/down, plus pitch, roll and yaw of the head position. In other words, the user can navigate the space, and might expect a subsequent shift in sound field both whilst moving and looking around. In contrast, 360º video (often presented as 'VR') does not permit locative movement within its space.

xvii 6DOF: Six Degrees of Freedom, which refers to movement in a 3-D-space: front/back, left/right, up/down, plus pitch, roll and yaw of the head position. In other words, the user can navigate the space, and might expect a subsequent shift in sound field both whilst moving and looking around. In contrast, 360º video (often presented as 'VR') does not permit locative movement within its space.

xviii A separate audio stem in the overall mix that does not respond to head tracking to provide elements such as narrative or beds.

xix A bed is an audio track that might be spatialized, but is not dependent on precise localization of any of its content. It might typically be a sonic foundation on which to place other more localized sources.

xx Also known as the mid-sagittal plane, this is the vertical plane that bisects the human body vertically through the naval. As such, strictly speaking it can represent perception of sonic height, but only when looking directly ahead.

xxi Also known as the mid-sagittal plane, this is the vertical plane that bisects the human body vertically through the naval. As such, strictly speaking it can represent perception of sonic height, but only when looking directly ahead.

xxii The HRTF actually exists in the frequency domain. The temporal equivalent is known as Head-Related Impulse Response (HRIR).

xxiii The HRTF actually exists in the frequency domain. The temporal equivalent is known as Head-Related Impulse Response (HRIR).

xxiv Such processing is performed by convolution, a CPU-intense mathematical operation commonly used in reverbs whereby every sample in the source is multiplied by every sample in the filter in order to impose a spectral fingerprint on the source.

xxv Such processing is performed by convolution, a CPU-intense mathematical operation commonly used in reverbs whereby every sample in the source is multiplied by every sample in the filter in order to impose a spectral fingerprint on the source.

xxvi Both fixed two-channel, and real-time multi-channel ambisonic that might also respond to head tracking

xxvii HRTFs are linked to in-ear or over-ear designs.

xxviii Typically: Pitch – analogous to nodding, Yaw – looking side to side, and Roll – moving the ear towards the shoulder.

xxix Typically: Pitch – analogous to nodding, Yaw – looking side to side, and Roll – moving the ear towards the shoulder.

xxx Other systems are also possible, which might spatialize individual tracks in middleware or a game engine.

xxxi Other systems are also possible, which might spatialize individual tracks in middleware or a game engine.

xxxii Head locking is where the sound is independent of movement, just as with normal headphone listening. A separate buss that bypasses the rotator is required for such parts.

xxxiii The process of converting one playback format to another of a greater number of channels.

xxxiv The process of converting one playback format to another of a greater number of channels.

xxxv In fact, it was shown that a cut-off frequency could be anywhere between 1 and 4 kHz for this effect to hold.

xxxvi In fact, it was shown that a cut-off frequency could be anywhere between 1 and 4 kHz for this effect to hold.

xxxvii To maintain something close to stereo theory, although this only holds when facing forward towards the speakers, and will break down for say a lateral placement.