The Why and How of With-Height Surround Sound

Jörn NETTINGSMEIER
Freelance audio engineer
Lortzingstr. 11
Essen, Germany, 45128
nettings@stackingdwarves.net

Abstract

With-height reproduction is a hot marketing item in surround sound. This paper examines the (sometimes non-obvious) motivations behind it and discusses the advantages and shortcomings of different methods as to the perceptional mechanisms of height localisation.

Keywords

With-height surround sound, Ambisonics, Psychoacoustics

1 A brief history of height reproduction

With-height surround has featured in many experimental one-off installations such as Stockhausen's "Kugelauditorium" at the 1970 World's Fair, or Bayle's "Acousmonium". They have usually employed custom-tailored, ad-hoc driving techniques with little or no regard for portability. The speaker system is considered an integral part of the artwork or performance rather than an interchangeable tool.

Therefore, they are only of historical interest today, although a number of sophisticated systems in the acousmatic tradition still exist and continue to be developed. On the other hand, there have been numerous proposals to bring with-height surround to a wider market in a systematic and portable way.

As early as 1973, Ambisonics pioneer Michael Gerzon suggested a practical approach to what he called *periphonic* sound using only four channels, also known as B-format [1]. In 1992, he proposed the technology as a candidate for the then-upcoming HDTV standard [2].

In 1999, Tomlinson Holman demonstrated "10.2", the first commercial cinema sound proposal to intclude height channels, if only at the left and right front [3].

Around the same time, German tonmeister Werner Dabringhaus entered the budding DVD audio market with his 2+2+2 system [4], which trades the center and LFE channels of a 5.1-capable medium for left and right frontal height.[1]

Belgian sound engineer Wilfried van Baelen experimented with 2+2+2 in 2005, and extended the concept into what he calls Auro-3D [5]. In its simplest form, it adds four height speakers on top of the standard 5.1 layout.

Likewise in 2005, a team of NHK researchers led by Kimio Hamasaki introduced 22.2 to accompany a future ultra-high-definition TV standard. It features a complete upper ring of eight channels plus one zenith speaker, ten on the equator, and an additional three bottom channels in the front [6].

2 Classification of existing methods

With the exception of Ambisonics, the approaches mentioned so far all build upon (or drag along, depending on your point of view) previous technology. They are *channel-based*, which means that the mix has to be made specifically for the reproduction speaker layout at hand, and they

[1]Omitting the center channel might seem strange today, but in 2000, dedicated LCR microphone techniques were not in common use, and classical Tonmeisters in particular were mostly unaware of their potential. Stereo main microphone techniques on the other hand were well understood and mastered, and those who just tried to add a center without changing their L/R miking as well found only increased coloration and loss of imaging clarity.

employ stereophonic localisation techniques to create phantom sources between speakers.

Among manufacturers of Wave Field Synthesis systems, the term “3D” has been (ab)used in marketing for a long time even in the absence of height capability. More recently however, WFS manufacturer IOSONO has introduced elevated speakers, thus earning the 3D moniker, while at the same time increasing the tweeter spacing of their systems considerably.[2] [7]

Such proprietary "hybrid WFS" systems use undisclosed panning techniques to include height, very likely a combination of delay panning and VBAP. The latter, short for Vector-base Amplitude Panning, is a conceptually very simple and elegant method introduced by Pulkki in 1997 [8]. It extends the idea of level panning to triplets of speakers, allowing the positioning of a sound source anywhere on a speaker mesh surface. VBAP can be applied to arbitrary speaker layouts, but it produces timbre shifts and highly variable perceived source width depending on source location.[3] The worst-case source width is equivalent to Ambisonics (which delivers perfectly constant panning). [9]

This effect is even more pronounced with sparse arrays, making it a less-than-ideal approach for layouts such as Auro-3D, and it does not work at all if a given channel is reproduced over multiple speakers, as is common in cinema installations.

WFS, VBAP and similar techniques are *object-based*, which means that individual (usually monophonic) audio files are stored with separate positional metadata, allowing them to be modified and re-positioned easily. Another advantage is that the mix is decoupled from the layout of the speaker system used for reproduction. On the downside, it is quite cumbersome to describe natural ambient recordings (which describe a spatial continuum rather than individual “spatial samples”) as monophonic objects.

Furthermore, object-based systems require an elaborate and CPU-intensive rendering process for listening, with a complexity growing linearly with the number of objects (O(N)).

Finally, Ambisonics is *soundfield-based*: the B-format carries an arbitrarily precise description of the resulting physical soundfield, where precision is determined by the order. It is not easily possible to separate single objects, but spatially continuous ambience recordings can be included perfectly. The B-format is again decoupled from the speaker layout by means of a decoder, which contains the information about the speaker positions. The decoding step is trivial compared to WFS rendering, and its complexity is constant (O(1)).

Crosstalk-cancelled binaural (or *ear-signal-based)* systems are theoretically able to deliver height cues as well. In a virtual environment at RWTH Aachen, striking effects have been demonstrated [10], but without head-tracking and individual HRTFs and in the absence of visual cues, the results will be mixed at best. In any case, the coloration is severe, and systems can accommodate at most one or two listeners.

Headphone binaural systems can deliver perfectly convincing height with excellent fidelity, but the required 3-axis head tracking systems are not yet widely available, and perfect results require individual HRTF measurements for each listener. Catering to more than one listener implies that the rendering and head tracking has to happen in the headphones, which is not feasible with current embeddable computing platforms, or that an individually pre-rendered signal is presented to each headphone, which puts a great strain on wireless bandwidth.

3 Elevation perception and stereophony

There are two totally distinct motivations for the inclusion of height speakers.

The most obvious one is the desire to position or reproduce sounds along the vertical or z-axis, not just on the horizontal plane around the listener. The

[2] This implies that textbook WFS only happens at very low frequencies, and other localisation mechanisms must be employed for the remainder of the spectrum.

[3] In its basic form, VBAP will employ one speaker if the source is directly on a speaker position, two if it is on the line between two speakers, and three anywhere else.

channel-based systems mentioned above perform poorly in this respect, because they rely on *stereophonic* localisation. Its mainstay is the clever delivery of artificial interaural level and time difference cues (ILD and ITD).

However, ILD and ITD remain constant as a source moves upwards on the median plane, which explains the comparatively poor vertical discrimination of the human hearing apparatus. The only height cue available to us is a rather subtle coloration of the sound caused by reflection, refraction, and absorption effects on pinnae, head, and torso. This cue is purely monaural.

Height perception is most acute when the subject has a clear mental reference of the natural (i.e. non-elevated) timbre of a sound source and will degrade with synthetic or otherwise unfamiliar sounds (compare [11]).

Blauert has demonstrated that height perception correlates with the spectral distribution of the sound event, provided it is sufficiently broad-band. [12]

With narrow-band signals, the location of the auditory event on the median plane depends only on the frequency, not the actual sound source location. [13]

Hence, height illusion can be created by applying equalisation which exploits these effects.

Even though sound engineers or electro-acoustic composers may sometimes employ such EQ ad-hoc, none of the with-height stereophonic methods include it as part of their standard in any systematic way.

The only remaining tool for positioning a source along the z-axis is simple amplitude panning. However, such a vertical “phantom source” will not result in any ILD or ITD information. Even worse, the coloration cue will very likely be meaningless, too, because the sum of a horizontal and an elevated pinna-colored sound is not necessarily similar to the pinna effect on a physical source in between.

Consequently, stereophonic systems exhibit a very steep localisation curve along the z-axis. They will usually produce auditory events either on the equatorial or the elevated speaker level. While vertical motion can be suggested, stationary sources between the two extremes are not stable. [14]

Producing stable auditory events above the elevated speakers is likewise impossible.

The only techniques which can deliver good localisation at arbitrary locations outside the equatorial plane are hybrid "WFS", VBAP, and higher-order Ambisonics, if and only if the speaker density is sufficient in the region of the desired auditory event.

Microphone arrays for stereophonic with-height systems will usually aim at complete decorrelation between the corresponding horizontal and elevated channels by using vertically widely spaced omnidirectional microphones, or they may try to minimize crosstalk by using highly directional ones. Both approaches clearly do not aim for a continuum of localisation along the z-axis.

Which leads to the second, dominant motivation for height speakers: timbre. Proponents of 2+2+2 and Auro-3D in particular claim that, in addition to a more convincing feeling of envelopment, the perceived tone color will be more natural in the presence of appropriate height signals. Furthermore, the listening area is believed to be larger than that of a comparable horizontal-only system.

In the author's listening experience, this is generally true, not only for Auro-3D but also for the "hybrid WFS" systems mentioned before, as well as for Ambisonics. Informal A/B tests (performed by comparing the full rig with the equatorial speakers only), suggest three advantages of with-height systems:

They appear to be more robust in the presence of room problems, perhaps because more room modes are being excited but at lower level, which might even out coloration effects.

Furthermore, they provide a more realistic and more stable sense of envelopment, which facilitates suspension-of-disbelief.

Finally, the height speakers seem to smooth the Ambisonic phasing artefacts and timbre shifts across the listening area.

4 Height reproduction in Ambisonics

In Ambisonic systems of sufficiently high order, a coherent sound field is being reconstituted in the sweet spot. While a listener may still have trouble

discerning or localizing sounds along the z-axis due to the limited resolution of the human hearing apparatus on the median plane, s/he will be able to resolve these ambiguities by moving the head.

Small subconscious movements can provide subtle differential directional cues, and intentional tilting of the head can be used to train the more acute lateral hearing mechanism on the source and gather additional information [15]. This helps the brain to fuse very stable auditory events at height, but it requires a natural soundfield, without any psychoacoustic tricks optimized to deliver artificial cues.

It appears that the source localisation remains stable even after the head is back in the rest position, as if the brain memorises the reference and uses it to align subsequent ambiguous cues.

Interestingly, despite the typical checkerboard interference pattern of Ambisonic systems, head movements seem to remain beneficial even for listeners well outside the sweet spot, an observation that needs to be validated by further study.

High-fidelity soundfield reproduction without assumptions as to listener orientation and temporal or spectral trickery will let listeners explore the spatial structure of a sound scene individually. They will find that they can rely employ their entire auditory sensorium to perceive and analyse selectively whatever sparks their interest. It is safe to assume that this helps to increase enjoyment and depth of understanding.

In contrast, height illusions created by naive pinna coloration simulation or non-tracked crosstalk cancelled binaural will fall apart the moment the head tilts away from the assumed frontal direction, failing the listener just at the time s/he displays particular interest.

5 Benefits of periphonic sound reproduction

The perceptory advantage of soundfield-reproducing techniques[4] opens up new applications which would be very hard or impossible to achieve with stereophonic with-height systems. Consider the recording and reproduction of works whose spatial organisation demands exact vertical localisation (think "spatial fidelity") rather than a vague notion of spaciousness, or of multi-layered compositions which are sonically so complex as to be downright indigestible unless the individual components are very precisely discriminated in space.

Periphonic soundfield reproduction in higher-order Ambisonics requires a substantial additional investment in both equipment and effort over conventional stereo, and it should be obvious that the justification of this investment depends on the program material.

A novice to baroque music may benefit from a little spatial separation when learning to appreciate a three-part invention or a five-part fugue from the Well-tempered clavier, but this is more a teaching aid than an artistic requirement. The musical structure itself is so resilient to spatial and timbral modifications (as aptly demonstrated by Wendy Carlos [16] and many others) that it will fare surprisingly well even as a phone ringtone.

A Gabrieli double choir from the same epoch however might become too dense if reproduced in stereo, and a historically informed reproduction would mandate at least horizontal surround.

Recordings of complex organ works can be more approachable if the original vertical separation of the divisions is retained.

Ambient nature recordings are an obvious case in point for full periphony.

Even if all direct sound emerges from a limited set of directions, full periphony is required to render the room acoustics correctly, if desired.

On the other end of the spectrum, a low-down blues song is adequately conveyed by a minimalist reproduction system; the impact of a distinctly low-fi bootleg might actually be harmed by uncalled-for technological “sophistication”. Buyer beware.[5]

[4] In theory, this includes WFS. However, "classical" WFS systems do not deal with height and "hybrid" systems are not wave field synthesis in the strict sense.

[5] In all things of technical hack value, this author considers « because we can » a perfectly adequate justification. Yet in the arts (or any other form of inter-individual communication), the very same approach can be disastrous.

6 Conclusion

With-height systems in general have the potential to be more robust than horizontal-only setups. Even if height localisation is not strictly necessary for the job at hand, the improved envelopment and timbral benefits are arguments in favour of full periphony. Stereophonic methods are of limited use for vertical localisation. Of the different approaches, those that attempt some degree of soundfield reconstruction should be more robust than ad-hoc, channel-based approaches.

[1] Michael A. Gerzon, 1973: "Periphony: With-Height Sound Reproduction", in: Journal of the Audio Engineering Society, vol. 21, pp. 2-10
[2] Michael A. Gerzon, 1992: “Hierarchical System of Surround Sound Transmission for HDTV”, in: Proceedings of the 92nd AES Convention, Vienna
[3] Barry Willis, 1999: “Holman Conducts First Public Demo of '10.2' Surround Sound”, in: Stereophile website, http://www.stereophile.com/news/10489/
[4] Dieter Steppuhn, 2001: “Begegnung anderer Art. Neue Audio-DVDs '2+2+2' bei MDG”, in: Neue Musikzeitung 02/01, http://www.nmz.de/artikel/begegnung-anderer-art
[5] Auro-3D product website, history section, http://www.auro-3d.com/about-history
[6] Kimio Hamasaki et al., 2007, “22.2 Multichannel Sound System for Ultra High-Definition TV ”, in: SMPTE Motion Imaging Journal vol. 117(3), pp *40-49*.
http://www.nhk.or.jp/digital/en/technical/pdf/IBC2007_08040907.pdf
[7] Frank Melchior, 2011: “IOSONO: 3D audio solutions ”, Workshop held at the International Conference on Spatial Audio, Detmold, Germany
[8] Ville Pulkki, 1997: “Virtual sound source positioning using vector base amplitude panning.”, in: JAES vol. 45(6), pp. 456-466.
[9] Franz Zotter et al., 2010: “Techniques and Considerations on Sound Field Recording and Reproduction Using Spherical Harmonics”, p.19.
http://iaem.at/Members/zotter/publications/IEM_SHTechniques_Zotter.pdf/view
[10]Tobias Lentz, 2006: “Dynamic Crosstalk Cancellation for Binaural Synthesis in Virtual Reality Environments”, in: JAES, vol. 54(4) pp. 283-294
[11]G. Plenge, und G.Brunschen, 1971: “Signalkenntnis und Richtungsbestimmung in der Medianebene bei Sprache”, in: Proceedings of the 7th International Congress on Acoustics, Budapest, 19 H 10, according to [12].
[12]Jens Blauert, 1997, 1983: “Spatial Hearing. The Psychophysics of Human Sound Localization”,
MIT Press, Cambridge, Mass., p. 109
[13]Jens Blauert, l.c., p. 45
[14]Günther Theile, and Helmut Wittek, 2011: “Principles in Surround Recording With Height”, in: Proceedings of the 130th AES Convention, London
[15] Jens Blauert, l.c., p. 181
[16]Wendy (then Walter) Carlos, 1968: “Switched-on Bach”, Columbia Records