Audio Engineering Society

Convention Paper 9869

Presented at the 143rd Convention

2017 October 18–21, New York, NY, USA

Development and application of a stereophonic multichannel recording technique for 3D Audio and VR

Helmut Wittek[1] and Günther Theile[2]

[1] *SCHOEPS Mikrofone GmbH, Spitalstr.20, 76227 Karlsruhe, Germany*

[2] *VDT, Germany*

Correspondence should be addressed to wittek@schoeps.de

ABSTRACT

A newly developed microphone arrangement is presented which aims at an optimal pickup of ambient sound for 3D Audio. The ORTF-3D is a discrete 8ch setup which can be routed to the channels of a 3D Stereo format such as Dolby Atmos or Auro3D. It is also ideally suited for immersive sound formats such as wavefield synthesis or VR/Binaural, as it creates a complex 3D ambience which can be mixed or binauralized.

The ORTF-3D setup was developed on the basis of stereophonic rules. It creates an optimal directional image in all directions as well as a high spatial sound quality due to highly uncorrelated signals in the diffuse sound. Reports from sound engineers affirm that it creates a highly immersive sound in a large listening area and still is compact and practical to use.

1 Introduction

Recording engineers who work with 3D sound face a difficult task when choosing a suitable recording technique. The number of channels is greater than with playback systems that operate only in the horizontal plane, so the complexity increases as well.

When a customer demands 3D Audio rather than conventional 5.1 surround it may be tempting to apply solutions that are overly simple. But when a 3D recording has been made well, using a suitable recording technique, the advantages are impressively audible.

What is 3D Audio?

The approaches included in "3D Audio" reproduce sound from all spatial directions [1][2].

This includes:

- **soundfield synthesis/reconstruction** approaches such as Ambisonics and wavefield synthesis systems;
- **binaural** / virtual reality ("VR") systems; and
- **stereophonic systems** such as Dolby Atmos and Auro3D

3D Audio can give distinctly better spatial perceptions than 5.1. Not only is the elevation of

sound sources reproduced, but noticeable improvements can also be achieved with regard to envelopment, naturalness, and accuracy of tone color. The listening area can also be greater; listeners can move more freely within the playback room without hearing the image collapse into the nearest loudspeaker.

Why is Stereo different?

It is crucial to differentiate between "soundfield reconstruction" and stereophonic techniques because they differ fundamentally in the principle by which sources are perceived, as found by Theile [3][4]. In contrast to the common theory of "summing localization," Theile assumes that loudspeaker signals are perceived independently, and that their level and time differences thus determine the location of phantom sources just as in natural hearing. It is essential that this superposition of only two loudspeakers does not lead to audible comb filtering, as the physical properties of the sound field would suggest. A stereophonic system can very easily create phantom sources in various directions, with good angular resolution and without sound-color artifacts. This makes it superior to imperfect soundfield reconstruction principles such as wavefield synthesis with excessive loudspeaker spacing, or Ambisonics of too low an order, both of which create artifacts [5].

When recording and reproducing stereophonically, closely-spaced microphone pairs are used, which create time and/or level differences between the microphone signals. These signals are routed discretely to the loudspeakers. The interchannel differences lead to the creation of phantom sources [6]. Stereophonic systems with more than two channels, such as 5.1 or 9.1 Surround, may be considered as systems consisting of multiple individual loudspeaker pairs with time and/or level differences that create phantom sources [2].

There is a fundamental difference between a first-order Ambisonics microphone and a stereophonic array for 5.1, even though the microphone arrays may look similar. An Ambisonics array aims for physical reconstruction of the original sound field, but cannot achieve it because of the early truncation of the order of the reproduced spherical harmonics. A stereophonic array aims to capture time and/or level differences in individual microphone pairs, but often cannot achieve that because of excessive crosstalk between the pairs. Hence both approaches have their own artifacts, as well as methods for overcoming them [5][1].

What is an ambience microphone?

Often the sound source to be recorded is a speaking voice, an instrument or the like. These sources can easily be recorded with a single microphone, and reproduced either by one loudspeaker or panned between two loudspeakers. If multiple individual sources have to be captured, e.g. a pop band with four instruments, multiple individual microphones can be used. However, if the sound source is spatially extended, if the room sound is to be captured as well, or if there simply are too many sound sources, this method fails. In that case a so-called "main microphone" or "room microphone" pair/setup serves for the stereophonic pickup of these sources in an efficient way, because these arrangements of two microphones (or the five microphones of a stereophonic array for 5.1 surround) are designed so that the recorded scene is properly reproduced between the loudspeakers [6]. Typical "main microphone" techniques are A/B, ORTF and X/Y (for two-channel stereo), and OCT, IRT Cross/ORTF Surround or a Decca Tree (for 5.1 surround).

An "ambience microphone" arrangement is a "main microphone" arrangement as well. The only difference is that the sound source is 360° around the listener instead of only in front (as in concert recording). Hence an ambience microphone has no "front" direction, but an equally-distributed image of phantom sources throughout the entire space spanned by the loudspeakers. Often the Center channel is omitted in the design of an ambience microphone, because it would destroy this equality of energy distribution.

One recording method for all 3D formats?

There are various 3D Audio playback systems, so the recording techniques that work best for each of

them will naturally be different. For soundfield synthesis systems, multichannel microphone arrays can be a solution, while for 3D stereo, stereophonic miking techniques are the norm. For binaural reproduction in the simplest case, a dummy head can be used.

But all these systems share one requirement when recording complex, spatially-extended sound sources such as ambient sound: stereophonic techniques must be used, because they alone offer both high-quality sound and high channel efficiency (even two channels may be enough). It is impossible or inefficient to reproduce in high quality the sound of a large chorus, for example, or the complex, ambient sound of a city street, by compiling single point sources recorded with separate microphones.

In the same way, multichannel microphone arrays for soundfield synthesis, such as higher-order Ambisonics ("HOA") or wavefield synthesis, fall short in practice because their channel efficiency or sonic quality are too low. If on the other hand the number of channels is reduced, *e.g.* with first-order Ambisonics, the spatial quality becomes burdened with compromise.

For binaural playback, the dummy head technique is clearly the simplest solution—but it does not, in itself, produce results compatible with virtual reality glasses, in which the binaural signals must respond to the user's head motions. That would be possible only through the “binauralization” [13] of a stereophonic array—a technique that is already well established.

Is first-order Ambisonics adequate for 3D?

There is a common assumption that Ambisonics would be the method of choice for 3D and VR. The professional recording engineer would do well to examine the situation more closely.

Ambisonics, which has existed for a long time by now, is a technology for representing and reproducing the sound field at a given point. But just as with wavefield synthesis, it functions only at a certain spatial resolution or "order". For this reason, we generally distinguish today between "first-order" Ambisonics and "higher-order" Ambisonics ("HOA").

First-order Ambisonics cannot achieve error-free audio reproduction, since the mathematics on which it is based are valid only for a listening space the size of a tennis ball. Thus, the laws of stereophony apply here—a microphone for first-order Ambisonics is nothing other than a coincident microphone with the well-known advantages (simplicity; small number of recording channels; flexibility) and disadvantages (very wide, imprecise phantom sound sources; deficient spatial quality) of that approach in general.

Creation of an Ambisonics studio microphone with high spatial resolution is an unsolved problem so far. Existing Ambisonics studio microphones are all first-order, so their resolution is just adequate for 5.1 surround but too low for 3D Audio. This becomes evident in their low interchannel signal separation as well as the insufficient quality of their reproduced spatiality.

The original first-order Ambisonics microphone was the Soundfield microphone. The Tetramic [7] or the Sennheiser Ambeo microphone have been built in a similar way. The Schoeps "Double M/S System" [8][9] works in similar fashion, but without the height channel.

Ambisonics is very well suited as a storage format for all kinds of spatial signals, but again, only if the order is high enough. A storage format with only four channels (first-order Ambisonics calls them W, X, Y, Z) makes a soup out of any 3D recording, since the mixdown to four channels destroys the signal separation of the 3D setup.

Ambisonics offers a simple, flexible storage and recording format for interactive 360° videos, *e.g.* on YouTube. In order to rotate the perspective, only the values of the Ambisonics variables need be adjusted. Together with the previously mentioned small first-order Ambisonics microphones, 360° videos are very easily made using small, portable cameras.

For virtual reality the situation is different, however. The acoustical background signal of a scene is generally produced by "binauralizing" the output of a virtual loudspeaker setup, *e.g.* a cube-shaped

arrangement of eight virtual loudspeakers. The signals for this setup are static; turning one's head should not cause the room to spin. Instead, head tracking causes the corresponding HRTFs to be dynamically exchanged, just as with any other audio object in the VR scene.

As a result, most of the advantages of first-order Ambisonics do not come into play in VR. On the contrary, its disadvantages (poor spatial quality, crosstalk among virtual loudspeaker signals) only become more prominent.

If practical conditions allow for a slightly larger microphone arrangement, an ORTF-3D setup would be optimal instead as an ambience microphone for VR.

2 Criteria for stereophonic arrays

Stereophonic arrays are thus the approach of choice for ambience recording in all 3D formats. The requirements for 3D are the same as in two- and five-channel stereophony [1]:

- Signal separation among all channels in order to avoid comb filtering: No one signal should be present at significant levels in more than two channels.
- Level and/or arrival time differences between adjacent channels to achieve the desired imaging characteristics
- Decorrelation of diffuse-field sound for optimal envelopment and sound quality

2-channel stereophony

These demands are still easy to fulfil in two-channel stereophony; a suitable arrangement of two microphones and two independent channels can provide the desired imaging curve. Tools such as the "Image Assistant" [4] application (available as an iOS app or on the Web at www.ima.schoeps.de) have been developed for this purpose.

They take into account not only the creation of phantom image sources, but also the ever-important channel decorrelation. A classic, positive example is the ORTF technique, which has a 100° recording angle and delivers a stereo signal with good channel decorrelation.

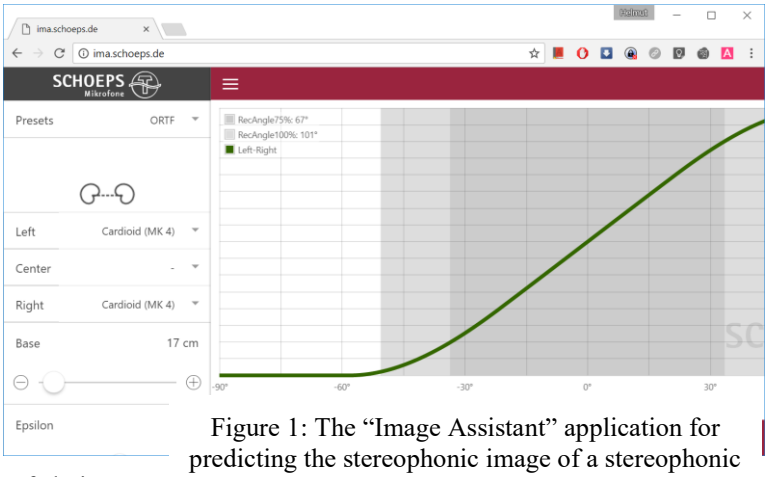

Figure 1: The "Image Assistant" application for predicting the stereophonic image of a stereophonic array (available as an iOS app or on the Web at www.ima.schoeps.de)

5-channel stereophony

The above requirements are distinctly more difficult to meet with five channels, and there are numerous geometries that fail to meet them, *e.g.* a microphone that looks like an egg the size of a rugby ball, with five omni capsules that can deliver only a mono signal at low frequencies.

Five independent channels simply cannot be obtained with any coincident arrangement of first-order microphones. A coincident arrangement such as first-order Ambisonics is thus a compromise for 5.1, though highly workable because of its advantages in compactness and post-production flexibility.

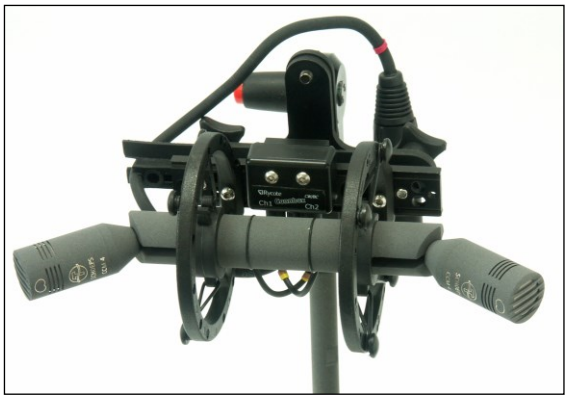

Figure 2: Two-channel ORTF system in a suspension designed for use within a windscreen; two cardioids, 17 cm, 110º

One optimal solution for ambient recordings in multichannel stereophony is the "ORTF surround" system, in which four supercardioids are arranged in a rectangle with 10 x 20 cm side lengths. Here the distances between microphones help with decorrelation, and thereby lend the sonic impression its spatial openness. The microphone signals are routed discretely to the L, R, LS and RS channels. The signal separation in terms of level is ca. 10 dB; thus, the sonic image during playback is stable even in off-axis listening positions.

Figure 3: Four-channel "ORTF Surround" system; four supercardioids, 10 / 20 cm spacing, 80º / 100º angles

8 or more channels

With eight or nine channels, the arrangement of the microphones becomes very difficult if the above-mentioned requirements are to be met. The simplest method for maintaining signal separation is to set up eight or nine microphones far apart from one another. Thus, a large nine-channel "Decca Tree" arrangement is very well suited for certain applications, although it has severe disadvantages that limit its practical usability. For one, the sheer size of the arrangement is greater than 2 meters in width and height. And the signal separation in terms of level difference is nearly zero; every signal is more or less available in all loudspeakers. Thus, this array can represent a beautiful, diffuse spaciousness, but stable directional reproduction isn't achieved beyond the "sweet spot." This can be helped by adding spot microphones.

3 The ORTF-3D recording method

An optimal ambience arrangement for eight channels is offered by the new "ORTF-3D" system developed by Wittek and Theile. It is more or less a doubling of the "ORTF Surround" system onto two planes, *i.e.* there are four supercardioids on each level (upper and lower), forming rectangles with 10 and 20 cm side lengths. The two "ORTF Surround" arrangements are placed directly on top of one another.

The microphones are furthermore tilted upward or downward in order to create signal separation in the vertical plane. Thus an 8-channel arrangement is formed, with imaging in the horizontal plane that somewhat corresponds to the "ORTF Surround" system. The microphone signals are discretely routed to four channels for the lower level (L, R, LS, RS), and four for the upper level (Lh, Rh, LSh and RSh).

In VR applications, virtual loudspeaker positions forming an equal-sided cube are binauralized.

Figure 4: A prototype of the ORTF-3D system at the ICSA conference in 2015. Eight supercardioids, horizontal distance 20 cm, vertical distance 0, angle 90º

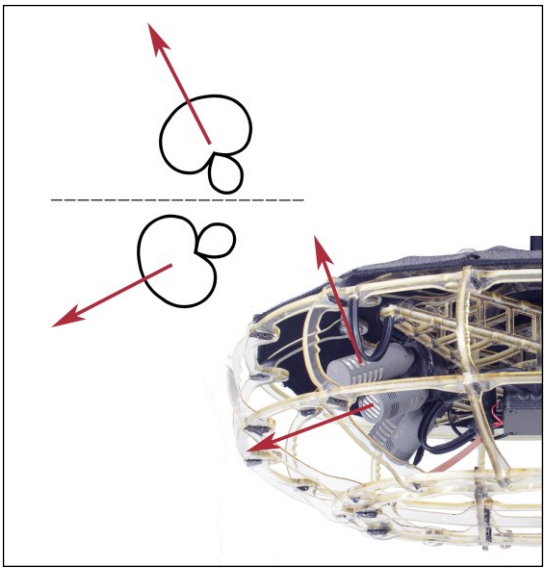

Figure 5: Orientation of the capsules: one vertical X/Y microphone pair for each vertical pair of loudspeakers

Lee et al. [11] found that the decorrelation of the diffuse field is less important in the vertical domain than in the horizontal domain. This means whereas it is clearly audible that an A/B microphone pair sounds wider than an X/Y pair when reproduced between L/R, there is only a little audible difference when reproduced between L/Lh. This helps a lot in the design of compact 3D ambience microphone.

Imaging in the vertical dimension is produced by angling the microphones into 90-degree X/Y pairs of supercardioids. Such a two-channel coincident arrangement is possible due to the high directivity of the supercardioids, and the imaging quality and diffuse-field decorrelation are both good.

This results in an eight-channel array with high signal separation, optimal diffuse-field correlation, and high stability within the playback space. All requirements are optimally fulfilled, yet the array is no larger than the compact ORTF Surround system—a decisive practical advantage.

Numerous test recordings have shown that the ORTF-3D approach produces very beautiful, spatially open and stable 3D recordings.

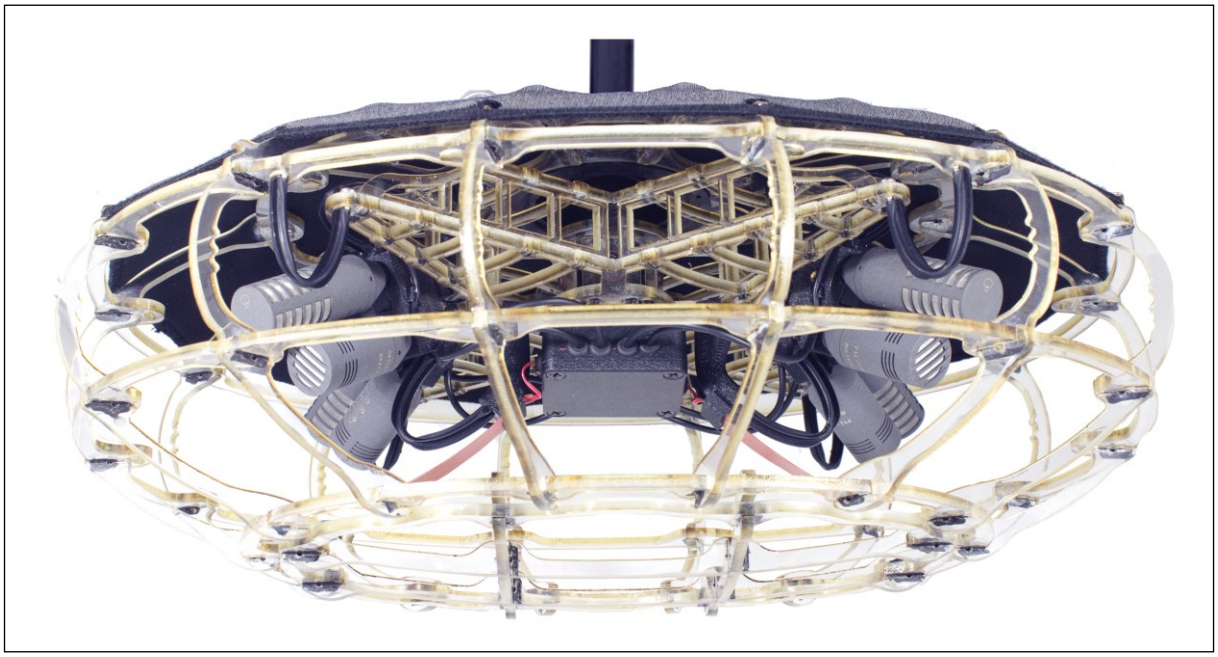

Figure 6: ORTF-3D arrangement, in a windscreen with the cover removed

4 Translating theory into practice

For the SCHOEPS ORTF-3D Outdoor Set [12][10], eight compact supercardioid CCM studio microphones are used. All microphones, as well as the windscreen itself, are elastically suspended in order to decouple vibrations. Each vertical X/Y pair is composed of one front-addressed CCM 41 and one radially-addressed CCM 41V. This enables a space-saving parallel arrangement of the microphone housings.

The windscreen and suspension have been developed by Schoeps together with the suspension specialist company CINELA. As with the "ORTF Surround" windscreen, elastic suspensions are also available for the ORTF-3D windscreen; fur, optional rain protection, multicore cables with breakout cables and integrated heating are standard. The windscreen is designed to be mounted by hanging. Long-lasting 24/7 outdoor installations, *e.g.* from the roof of a stadium, are possible.

This microphone arrangement, which was initially introduced as a prototype at the end of 2015, has already been sold or rented in considerable numbers to customers in the sports and VR sectors.

Tests have been made with great success during the past two years, including several well-known sporting events. Further test recordings are available for download from the Schoeps website [12].

Figure 7: Windscreen with synthetic fur covering or rain protection, plus integrated heating, for outdoor applications

5 Conversion of the ORTF-3D setup for Dolby Atmos and Auro3D

The eight channels of the ORTF-3D are L, R, LS, RS for the lower level, and Lh, Rh, LSh and RSh for the upper level. They are routed to eight discrete playback channels without matrixing.

The Center channel remains unoccupied. A Center channel is seldom desired in ambience recording; it would distort the energy balance between front and rear, and require significantly greater distances among microphones in order to maintain the necessary signal separation. If a Center signal should be necessary for a specific reason, *e.g.* to cover the shutoff of a reporter's microphone, a simple downmix of the L and R signals at low level is sufficient.

In Auro3D the loudspeaker channels L, R, LS, RS, HL, HR, HLS and HRS are fed.

With Dolby, the integration in the Atmos production environment is equally simple; the channels L, R, LS, RS are simply laid down in the corresponding channels of the surround level, the so-called "Atmos bed," whereas the four upper channels are placed as static objects in the four upper corners of the Cartesian space in the Atmos panning tool. These are then rendered in playback through the corresponding front or rear loudspeakers.

The below screen capture from ProTools, with the four Atmos panners as well as the monitoring application, illustrates this.

6 Conversion for VR

In a virtual reality ("VR") environment, 3D video and binaural sound are reproduced via VR glasses with headphones. Head position and rotation are processed in real time. 360° videos can also contain binaural sound, but only head rotation is processed, not the head position.

If binaural sound is to respond to head tracking, a dummy head cannot be used as the recording method since it allows only for one head angle. Instead, the following sound components are gathered separately and assembled:

- "Audio object" with dry sound
- Binaural (+ Room) filters: "HRTF" or "BRIR"

Usually the audio object, *e.g.* a character in a VR video game, is a single source with a certain distance and 3D direction. It consists of dry sound, which is then processed via binaural and room filters

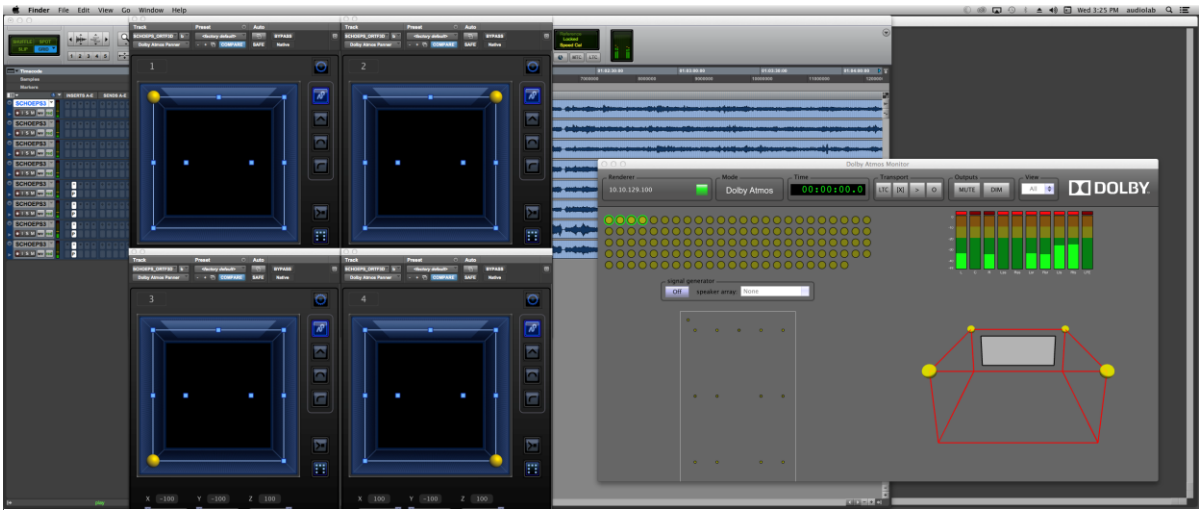

Figure 8: Routing of the eight channels from the ORTF-3D in Dolby Atmos (ProTools plugin)

(="binauralized") depending on its 3D direction. This direction is determined by the position of the audio object and the position and head rotation of the listener within the VR scene.

Figure 9: VR glasses (Samsung)

Current research [13] covers the individualization of HRTF filters: each listener could potentially choose an HRTF filter that corresponds to their own head/pinna/torso geometry and thus sound color artifacts and front/back-confusions can be avoided.

The acoustical background signal of a scene, or "ambience/atmo", is a very special kind of audio source. It cannot be recorded dry, nor can it be mapped to a single point source. In principle, it could be produced by the superposition of numerous audio sources in space, but often this would either be inefficient (*e.g.* trees in a forest) or impossible (live ambience from a venue).

Thus, a group of several audio objects forming an array of virtual loudspeakers is used to reproduce a stereophonic recording of the ambience. These group of loudspeakers can be chosen from a 3D preset, for example the Dolby setup 5.1.4, or the Auro3D setup 9.1, in each case without a Center loudspeaker. If no preset is available, one can define an equal-sided cube around the listener.

These audio objects are "diegetic" (=belonging to the picture), *i.e.* just like their visual counterparts, they do not move in response to head rotation. This does imply that their incidence angle in relation to the head changes with head rotations and thus the HRTFs change. “Non-diegetic” sounds are static and don’t change with head rotations, *e.g.* the voice of the narrator or accompanying background music. Their HRTFs stay stable with head rotations which means that the sound objects move relative to the picture!

The eight signals of the ORTF-3D microphone are reproduced on the group of 8 virtual loudspeakers to build up an optimal 3D live ambience in the VR environment.

First-order Ambisonic microphone for VR?

The use of a first-order Ambisonic microphone for this purpose cannot be recommended as described above. Being a small, coincident setup, its output lacks sufficient separation among channels, thus reducing the quality of its spatiality and 3D stereophonic imaging. A first-order Ambisonic microphone does have an advantage if the head rotation (or the movement of the angle of sight) is rendered by a change of the Ambisonics parameters – then HRTFs can be stable while the virtual loudspeakers are then “non-diegetic”. This can save performance of the system. In reality, this is seldom the case, as anyway the single audio objects are usually diegetic. Furthermore, many binaural renderers solve this problem by internally rendering a fine grid (*e.g.* Ambisonics 3rd order) of “non-diegetic” virtual loudspeakers on which the movements of the diegetic signals are simply routed by panning.

References

[1] Theile, G. and Wittek, H.: “Principles in Surround Recordings with Height”, 130th AES Convention, London, Mai 2011, Preprint No. 8403

[2] Theile, G. and Wittek, H.: “3D Audio Natural Recording (English, Natürliche Aufnahmen im 3D Audio Format)”, 27.Tonmeistertagung, Köln, November 2012

[3] G. Theile, “Über die Lokalisation im überlagerten Schallfeld” (“On Localization in the Superimposed Sound Field”), Ph.D. dissertation, Technische Universita¨t Berlin, Germany (1980)

[4] G. Theile, “On the Naturalness of Two-

Channel Stereo Sound," *J. Audio Eng. Soc.,* vol. 39, pp. 761–767 (1991 Oct.).

[5] Wittek, H., Rumsey, F. and Theile, G., "Perceptual Enhancement of Wavefield Synthesis by Stereophonic Means.", Journal of the AES, Volume 55 Number 9, September 2007

[6] Wittek, H. and Theile, G.: "The Recording Angle - based on localization curves", AES_112th_Convention_Paper (English), 2002

[7] CoreSound TetraMic: http://www.core-sound.com/TetraMic/1.php

[8] Wittek, Haut, Keinath: "Double M/S – a Surround recording technique put to test", 24. Tonmeistertagung 2006

[9] Schoeps Double M/S Systems: http://www.schoeps.de/en/products/categories/doublems

[10] ORTF-3D Article: http://www.hauptmikrofon.de/stereo-3d/3D Audio/ortf-3d

[11] H. Lee and C. Gribben, "Effect of Vertical Microphone Layer Spacing for a 3D Microphone Array," J. Audio Eng. Soc., vol. 62 (12), pp. 870-884. (2014 Dec.)

[12] SCHOEPS ORTF-3D Microphone: www.schoeps.de/ortf3d

[13] Rozenn Nicol, Laetitia Gros, Cathy Colomes, Markus Noisternig, Olivier Warusfel, et al.. "A Roadmap for Assessing the Quality of Experience of 3D Audio Binaural Rendering.", *EAA Joint Symposium on Auralization and Ambisonics*, Apr 2014, Berlin, pp.100-106, 2014.