3D Audio Natural Recording [1]

(Natürliche Aufnahmen im 3D-Audio Format)

Günther Theile, Helmut Wittek***

* VDT, theile@tonmeister.de

** SCHOEPS Mikrofone GmbH, wittek@schoeps.de

www.hauptmikrofon.de, Forum für Stereofonie- und Aufnahmetechnik

Abstract

New multichannel sound formats extending 5.1 with height channels are adding the third dimension to recordings. The height layer for the ITU 5.1 surround standard according to Auro 9.1 is providing a much wider range of spatial sound effects and more realism of spatial reproduction. It ensures a clearer representation of spatial depth, a more natural spatial impression, improved envelopment and thus immersive sensation. Prerequisite is the use of adequate microphone, mixing and rendering techniques. The psychoacoustic principles of stereophonic channel-based recording methods are discussed. Concrete proposals for microphone configurations can evolve from these considerations.

1. Introduction

After the international ITU-R BS.775-1 standard had been released in 1992, it took key-media vendors some time to implement the necessary techniques and to gain sufficient expertise in using them. In recording, switching from 2.0 to 5.1 was the first considerable step away from "pure" stereophony with two loudspeakers placed in front of the listener towards the realistic reproduction of an acoustic environment.

5.1, however, was just a compromise. It was necessary due to restrictions such as compatibility with 2.0 stereo and the fact that at that time cinema formats supported a maximum of six channels. Therefore, 5.1 essentially brought along not more but two improvements [1]:

- It increased the listening area and improved the stability and quality of stereo sound by subdividing the L/R basis, which is 60° in width, into two stereo sub-ranges with 30° each (L/C and C/R).
- Within certain limits, it allowed for creating a realistic acoustic environment by placing to additional surround speakers behind and on the sides of the listener.

A few years ago, we found that virtually the entire industry was ready for using 5.1 in production, distribution, and on end-user equipment. In addition, consumers today typically accept the presence of a larger number of speakers – at least when used as components of a home-cinema setup. On the other hand, we discovered that only a limited number of

/[1] Enhanced manuscript of AES Preprint 8403, 130th AES Convention 2011

listeners are able to achieve the sound quality that can actually be realized using a surround system – or the quality they had hoped for. There are several reasons for this:

- The listening environment is unfavorable in terms of room geometry or acoustics, the arrangement of the speakers is not standards-compliant, or device settings are inappropriate.
- The recording quality is bad. This results either from economic constraints in production or from inappropriately chosen miking and mixing techniques.
- The 5.1 listening zone is too narrow. There are recordings that require a perfect listener placement, assuming that only the sweetspot matters.
- Limitations of the 5.1 format including improper 3D imaging, proper speaker positioning in height and in relation to the listener's head, and imperfect distance imaging.

The above list is not necessarily in order of importance; however, it illustrates that problems arise mostly when it comes to practical application. This is equally true for the producer and the listener. Eliminating the issues just by increasing the number of channels and speakers is not possible; in fact, recently introduced enhancements and innovative systems ranging from various 7.1 formats to high-order ambisonics (HOA) and wave-field synthesis (WFS) require new paradigms, new hardware, and special attention from recording engineers. Plus the listener still needs to accept a living room in home-cinema style. In this context, the current variety of formats and the lack of standards present an additional obstacle. The current DCI specification (or SMPTE 428M, respectively) specifies channel mapping and purposely allows for any use of 16 channels.

The ITU-R BS.775-1 standard already specified optional LL and RR speakers located between the front and surround speakers. This improves the stereo quality of side imaging, enlarges the listening zone, and fills the gap between frontal and side imaging. Altogether, this leads to more flexibility for reproducing stationary audio events at the side or the critical lateral reflections. In conjunction with new developments in film sound, companies such as DTS and Dolby follow this principle and promote various 7.1 formats. These use a similar array where four surround speakers are spread laterally and behind the listening zone while utilizing the same front-speakers arrangement (L/C/R). Today, several hundred Blu-ray discs offering 7.1 audio are available for home-cinema use. Those media excel with clear sound definition and stable directional imaging at the sides and behind the listener; however, there are hardly any music recordings [2].

All those surround formats are essentially based on stereophony, i.e. they use phantom sources between two adjacent speakers for source imaging. In surround, the direction of the phantom source greatly depends on the listening position and is highly unstable; therefore, directional imaging virtually relies on the physical speaker positions. The volume balances are position-dependent as well. This is particularly true for the relation between front and surround sources. Therefore, adding more channels on the horizontal plane aims at enlarging the listening zone and providing a more homogenous and more stable directional resolution.

There are alternative ways of using additional channels, leaving the horizontal plane. Arranging speakers above the listener's head complements the spatial area, allowing for creating a 3D sound within certain limits. Almost ten years ago, *Werner Dabringhaus* published the first music recordings produced using his 2+2+2-recording technique. This approach is based on 5.1 but does without center and subwoofer speakers; instead, it uses two speakers positioned on top of L and R [3]. This concept was designed with the audio DVD in mind. The objective was to reproduce the sound from the concert hall as

realistically as possible, so it used speakers allowing for imaging height information rather than center and subwoofer speakers. Similarly, *Tom Holman* integrated the third dimension using two tilted height speakers placed in front of the listener; however, his 10.2 Channel Surround Sound setup requires eight channels on the horizontal plane and was originally created for cinema and home-cinema applications [4].

In 2006, *Wilfried Van Baelen* introduced the Auro 9.1 format that specifies four extra channels for height information. With the Auro 9.1 basic version, the height speakers complement the 5.1 format – they are positioned above the L, R, RH, and LH speakers (figure 1). Of course, similar formats such as 7.1 Surround can be complemented using four height speakers, for example, in a "quadraphonic" array. At last year's Tonmeistertagung, *Wilfried Van Baelen* lectured on the latest developments and experiences during the Digital Cinema session [5].

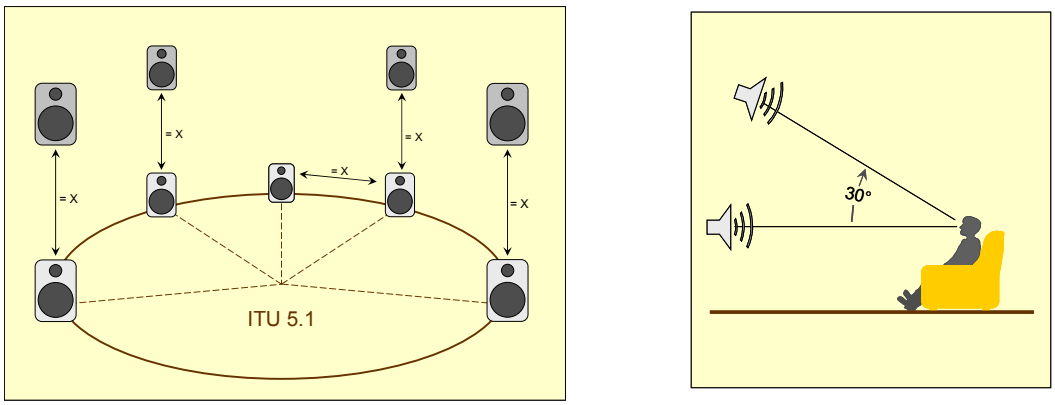


Figure 1:
Auro 9.1 basic setup (according to [5]), backward compatible with ITU-R BS.775-1

The main feature of this format is the cube-like arrangement of eight speakers. It allows for including the entire upper half space for the reproduction of (early) reflections and for appropriately reproducing the subjective spatial diffusion of the reverb part. The format has provided an excellent starting position for imaging parameters such as envelopment, spatial impression, and depth. In addition, the height speakers obviously offer the same possibilities for stereo imaging as the ITU setup without the center speaker. On the other hand, creating phantom sources between the lower and upper speakers (i.e. stable directions for stationary audio events with an elevation of 0 to 30°) as well as immediately above the listener's head is practically not possible. We will discuss this shortly.

Some limitations of the 5.1 format can be eliminated or alleviated using Auro 9.1; others cannot. Table 1 lists a number of attributes of reproduced sound. The first four parameters affect the direct portion (and are normally modified using panning); the other four attributes refer to the effects of indirect sound (designed using miking techniques and processing). These attributes allow for categorizing and comparing the profiles of the various techniques in a reasonably adequate fashion, provided that the reproduction recommendations have been implemented properly and appropriate miking and mixing techniques have been used on the recording side.

As the table shows, Auro 9.1 offers some specific benefits compared to other speaker-arrangement techniques. This also applies to other formats complementing 2D surround systems with quadraphonic speaker arrays on the plane above the listener. In section 4, we will describe arrangement options and limitations in detail with a focus on relevant miking techniques.

2. 3D Sound for 3D Video

New developments in dummy-head recording (e.g. [6], [7]) marked the start of serious and partly successful efforts to establish 3D in broadcast and on recording media. The original method of 3D audio is a binaural reproduction of ear signals. Ideally, the reproduced dummy-head signals are identical with the ear signals the listener would perceive at the dummy-head position inside the recording room. In this case, the virtual listening experience would match the real sound inside the recording room. Unfortunately, binaural techniques are limited to special applications due to various practical reasons [8]. They are not compatible with speaker reproduction, that is, their 2-channel signals cannot be converted to multichannel speaker signals producing the same effect. On the other hand, the quality of 3D imaging that can be achieved using binaural techniques may be used as a reference: The imaging zone includes the entire upper half space, and audio events of any elevation and distance can be represented.

In the interest of completeness, we also want to look at the intra-active perspective. This is a feature of natural auditory scene analysis. The way directions are perceived changes depending on the distance to the sources: when nearby sources move, they travel "further" than remote sources. WFS systems allow for reproducing this behavior [9] – a fact of that may be interesting, for example, for future developments in gaming. However, we will not go into the details in this paper.

2.1. Speaker Reproduction

One important point we want to focus on here is which features of 3D audio reproduction are appropriate for 3D video. The first thing we noticed is that the initial situation is different compared to audio. The flat two-dimensional image is converted to 3D video by creating a sense of depth using the means of stereoscopy within the limits of video reproduction[2]. Unlike, with audio, the third dimension is height (the other two being direction and distance). Regardless of the extent to which the possibilities of imaging are limited, 2.0 stereo, 5.1 surround, and WFS are definitely 2D techniques. This is particularly obvious with 2.0 stereo, which emulates distance and depth and limits the listening zone to a 60° angle; with 5.1 surround and WFS, the limitations are not so clear [1], [9] (see table 1).

A 3D video image can be conceived as a window to a three-dimensional scene with objects that may extend into that window. WFS basically allows for imaging the correspondence auditory objects in front of the speaker display. In this respect, to create the perfect match between audio and video by any means, one would have to utilize WFS and binaural techniques; however, this approach would be hardly worth the effort. There are several

[2] More precisely, a distinction is made between 2½-D reproduction (where the viewer moves to perceive depth) and 3D reproduction (depth is intuitively perceived due to stereoscopy).

reasons for this including the fact that WFS ignores the height attribute and appropriate reproduction of binaural signals to multiple listeners requires the use of headphones.

ATTRIBUTES OF SOUND REPRODUCTION	2.0 STEREO	5.1 SURROUND	AURO 9.1	WFS*	BINAURAL TECHNIQUES
Front direction	●	●●	●●	●●	●
Surround direction		●	●	●●	●●
Elevation			(●)***		●●
Height			●		●●
Distance/depth	(●)**	●	●●	●●	●●
Proximity to the head				●	●●
Intra-active perspective				●●	
Spatial impression	(●)**	●	●●	●	●●
Envelopment		●	●●	●	●●
Timbre	●●	●●	●●	●	●●

Table 1: Comparison of stereo/surround-format profiles
(requires appropriate recording and reproduction techniques)
*horizontal arrays; **emulated depth/spatial impression; ***unstable; on the sweetspot only

Therefore, Auro 9.1 (and above) is the format of choice. It meets many of today's requirements to a universal and compatible future-oriented standard for digital cinema, games, broadcast, and the music industry [5]. As we will describe in detail, engineers recording for an Auro 9.1 speaker array need to pay special attention to the phenomena of psychoacoustics in order to achieve good results when implementing specific creative ideas. After the introduction of the 5.1 surround channels, the inclusion of height has been the second step towards enhancing freedom in speaker stereophony. One of the most sophisticated tasks is recording music "realistically". It requires the use of a special miking technique to control the four main attributes of 3D recording at the same time – source direction and width, depth of the scene, spatial impression, and envelopment. Based on that recording situation, we will explore the new creative possibilities in the following sections.

2.2. Headphone Reproduction

Current convolution methods allow for realistically imaging a virtual Auro 9.1 studio using headphones. Commercially available Binaural Room Synthesis (BRS) systems ensure virtual 5.0 speaker reproduction in professional quality. In addition, they can easily be modified to support additional height channels. A BRS system convolves surround signals with the sampled binaural impulse responses (IRs) of a high-quality studio. Data suitable for convolution are selected using head tracking. This method takes the current head orientation into account, so the listener locates the virtual speakers regardless of the head posture (i.e. in relation to space) [10]. In 2007, the IRT released a BRS plug-in for VST-compliant host applications [11]. In the meantime, a cost-effective BRS standalone device capable of perfectly emulating the studio environment using individual equalization is available [12].

This technology allows for autonomously producing Auro 9.1 recordings on the OB truck and in any other scenario with unfavorable monitoring conditions. Engineers can take their familiar monitoring environments wherever they go. Several monitoring scenarios are available at the press of a key, allowing, for example, for checking the sound beyond the sweetspot or comparing various speakers or listening rooms. Using BRS, consumers can achieve significantly better reproduction quality with Auro 9.1 signals than living-room speakers would allow at all. In addition, BRS makes the listener completely independent from the selected speaker array: If fed with suitable material, a BRS processor can essentially emulate virtually any multichannel speaker setup. This eliminates all the practical problems that come up when placing speakers at home properly.

The BRS technology will considerably speed up the acceptance of production quality, multichannel audio, and, in particular, Auro 9.1 in the market.

3. Psychoacoustic Requirements to Multichannel Audio

The human ear evaluates various properties of the sound field and uses them for spatial hearing. Table 2 roughly outlines the meanings of direct sound, early reflections, reverb, and listener envelopment for each of the above sound attributes and the timbre. Enveloping sound includes both diffuse-field sound (background noise, "atmo") and audibly decaying reverb.

SOUND ATTRIBUTES IN THE HALL	DIRECT SOUND	EARLY REFLECTIONS	REVERB	BACKGROUND NOISE
Direction/elevation	●●	●		
Distance/depth		●●		
Spatial impression		●●	●	
Envelopment			●	●●
Timbre	●●	●	●●	

Table 2: Interrelation between sound attributes and sound-field types

The ear is typically capable of intuitively (or spontaneously) distinguishing between these three portions in natural sound; however, the more localization and timing are deteriorated due to inappropriate reproduction, the more difficult it is to achieve this intuitive distinction. A good example is a mono recording where direct sound, early reflections, and reverb sum up to a heavily colorized sound mush. In this case, spatial perception is exclusively based on conscious recognition. For example, a long reverb implies a large room, low-level direct sound means "long distance", etc.

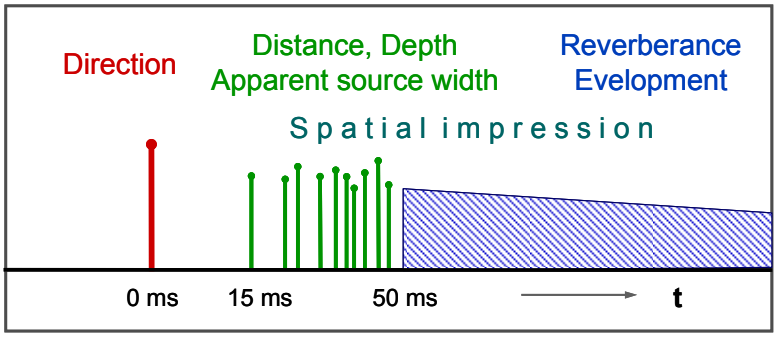


Figure 2: Influence of attributes on sound impression over time

3.1. Reflections and Reverb

Indirect sound portions allow for reproducing the recording space. The relation between direct and indirect sound determines the spatial attributes of a sound event. Figure 2 shows this interrelation. The natural pattern of early reflections occurring at a delay of 15 to 50 milliseconds plays a key role in spatial hearing. When it comes to recording, this portion of reflected sound deserves special attention as it critically affects attributes such as distance, depth, and spatial impression. The hearing takes spatial information from early reflections and converts it to a spatial event. With natural sound, the human ear performs this conversion spontaneously and with amazing robustness because that type of sound contains all properties of a reflection pattern in their original form. Key parameters include

- the timing structure in relation to direct sound
- levels and spectrums
- horizontal and vertical incidence directions

Imaging a spatial environment is realistic when the ear is able to recognize and interpret the features of the reflected sound – that is, when it "understands" the reflection pattern. Therefore, the reproduction must be absolutely consistent with a real spatial environment. The same applies to the spatial distribution of the early reflections' incidence directions. Meeting this requirement using room microphones is hardly possible (see section 4) because one needs to keep acoustic crosstalk on the ambient-microphone channels as low as possible (approx. 10 dB at the most). A single reflection coming in from a specific direction – say, the top-right corner of the rear part of the hall – should be reproduced as such; it must not be picked up by the "wrong" mikes. This would be the case, for example, when using omni directive microphones in a room-microphone array.

The perception of distances and depth mostly depends on early reflections. This can be proven by simply adding pure early reflections (without the reverb) derived from a real room to a source that has been dryly recorded. The source is perceived as remote, which is in correspondence with the reflection pattern. Perception is particularly stable when the reflections come in from the original directions of the upper half space. Reproducing depth requires careful handling of early reflections.

Adding appropriate reverb at a suitable level creates a natural sense of depth and realistic spatial impression /[3]. Even with short reverb times, the virtual reproduction of these two attributes creates a realistic spatial impression. Increasing the reverb time, for example, by using concert-hall or church reverbs, adds another attribute of spatial hearing: the envelopment.

3.2. Diffuse Background Noise

Background sound (or noise) consists of a large number of spatially distributed individual acoustic sources that cannot be separately localized. Rustling leaves in a wood, audience noise and response, and applause at a performance are typical examples. Unlike indirect sound, this portion of the surrounding sound cannot be created using effect units, so appropriate miking is essential.

When recording indoors, using room/ambience microphones for recording background noise as well is obvious. With some trying and testing, an experienced engineer can create a realistic balance (for example, upper/lower space) between reverb and background noise (applause, audience noise) by carefully selecting capsule and polar patterns and sensibly placing the microphones; however, there are situations where this cannot be achieved, and it should be avoided while actually recording. The use of an 8-channel reverb unit provides more flexibility: It allows for routing the background noise to the lower speakers while reverb is fed to all eight channels.

4. 3D Audio with Auro 9.1

The speakers on the upper plane obviously have the same imaging capabilities as those on the horizontal plane (except for the center speaker). The stereo image in the L/C/R range is complemented by 2-channel stereo sound on the upper L_h/R_h base. Similarly, the additional height speakers can be used in the same way as those on the horizontal plane. This arrangement already enhances flexibility considerably. An interesting aspect is the interaction of the two planes and its resulting possibilities. In the following sections, we will describe source imaging using the five speakers in front of the listener and the reproduction of reflections and diffuse sound field in the 3D surround array.

4.1. The Upper and Lower Representation Areas

Elevating Sources

Unfortunately, the familiar stereo imaging of localizable sources can be achieved only at the upper and lower edges of the area in front of the listener (i.e. between L-R and L_h-R_h). A localization of phantom sources between the upper and lower speakers is highly unstable due to propagation-delay differences and also depends on the spectrum. Elevation cannot be achieved just by using panning functions – this would affect sound and spatial perception in a way that cannot be controlled. Figure 3 shows a practical analysis of stereo-level relations

/ [3] The term “spatial impression” refers to the effect of early reflections and early reverb on localization. Due to reverberation inside the room, the apparent source width (ASW) seems greater, and the source event appears to be “fuzzy” in time.

between speakers arranged one above the other (0° and 45°) in front of the listener [16]. It is obvious that reliable localization cannot be achieved even from the sweetspot and with correct delay relationships; this is similar to lateral phantom sources. Thus, stationary-source elevation cannot practically be accomplished.

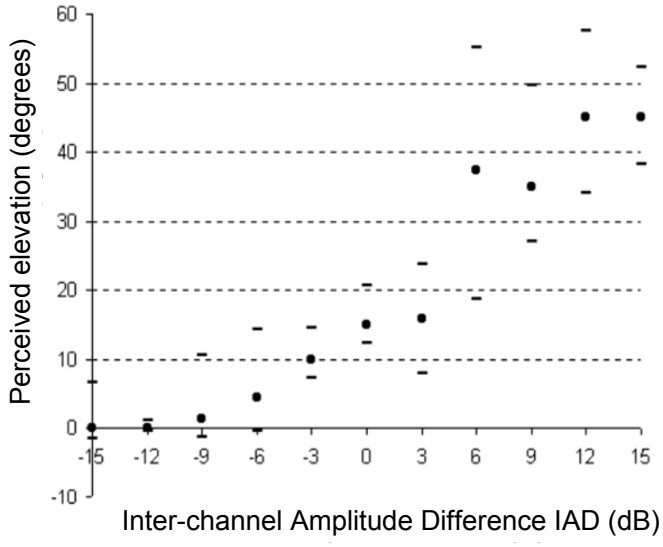


Figure 3: Stereo imaging on the median plane affected by differences in level (speaker angles: 0° and 45°), taken from [16].

In addition, very small differences in propagation delay result in the phantom source migrating upwards or downwards. A delay of just 0.5 milliseconds is sufficient to move the audio event to one or the other side. Coloration will occur as well. Thus, the listening zone is greatly limited regarding depth and height. Figure 4 shows the delay conditions in an Auro 9.1 home-cinema speaker array.

Elevating or upward-expanding a stationary source using the upper speakers is practically not achievable. This is particularly true where a large listening zone is required. Trying to solve this issue using panning functions would not be successful and would also result in coloration (which would, however, be masked almost completely by the diffuse-field portion). This scenario is similar to using the L/LS and R/RS side-speaker pairs – the speakers are the only stable source positions. Moving sources can, however, be represented within certain limits.

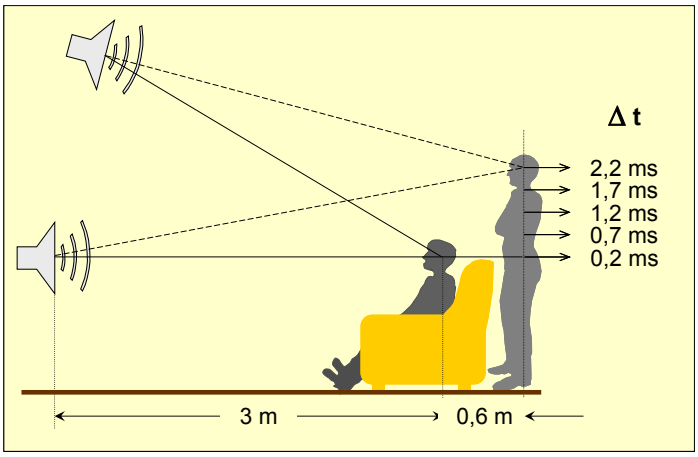


Figure 4: Delay differences occurring in listening positions beyond the sweet spot

Filling Up the Areas

Much better conditions exist for reproducing a large number of spatially distributed individual acoustic sources that cannot be separately localized. They have properties similar to those of a largely spaced A/B setup or a Decca tree: While directional imaging is not practicable due to the mapping curves being much too steep [8], reproducing a well-balanced imaging, for example, of a large orchestra and the reflections produced by it is possible. The risk of creating a "hole" in the center is controllable in many recording scenarios, in particular, where the diffuse-field portion dominates the sound. Therefore, filling the areas in height is actually possible and an important creative element.

4.2. Reflections and Diffuse Sound

The approach allows for distributing, in particular, the early reflections in the upper plane. This is due to the delay differences of individual reflections on the capsules. Reflections come in naturally from upper directions, too.

The preferable distribution of the reflections reduces their spatial density, allowing the ear to better distinguish spatial information. Figure 5 shows the effect for the transition from 2.0 to 5.1 to Auro 9.1. Another critical factor in this context is a positive effect on the timbre, which results in improved perception of reflections.

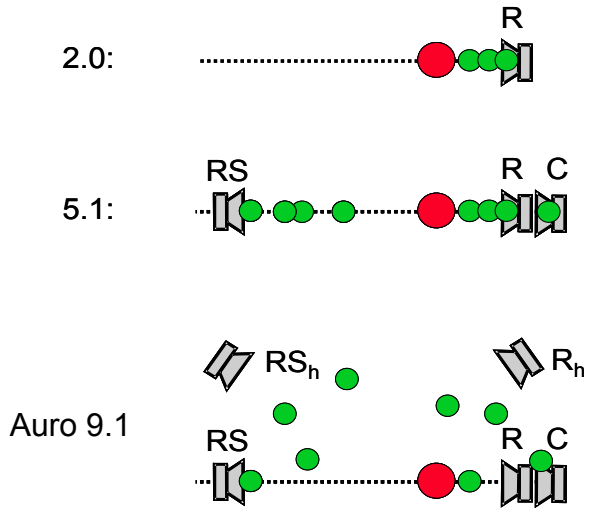


Figure 5: Spatial distribution of reflection patterns in 2.0, 5.1, and Auro 9.1

5. Recording for Auro 9.1

When looking for a suitable recording technique for Auro 9.1, knowing the principles of natural hearing is helpful. Considering the complexity of the subject, one may decide that trying and testing would be appropriate – after all, it rarely sounds worse when feeding any portion to the height speakers. However, it soon becomes clear that this is not what we want. We know from our experience in searching the proper recording technique for 5.1 that a discrete recording needs to be considerably better than an automatic upmix with respect to spatial depth, spatial impression, envelopment and even timbre.

One should refrain from trying to compare 5.1 and Auro 9.1 by just switching the height speakers off and on or creating a downmix. Such a comparison would be misleading. The listener/consumer needs to be convinced of the true added value offered by the individual professional performance that goes along with this innovative reproduction technique. For that purpose, we need not only to improve spatial reproduction but also require new ideas for an aesthetic use of the height channels.

The purpose plays a key role in determining the suitable recording technique. There are techniques that are more suitable for delivering convincing spatial imaging, and others that are better for use with spot microphones. General guidelines may be found on the basis of great recordings. However, psychoacoustic expertise and practical investigations are essential to be able to verify, refine, and adapt them for individual given recording situations and aesthetical intentions.

5.1. Channel Separation

To create the spatial resolution of direct sound, background / atmo , and/or early-reflection portions as described above, microphone placement needs to ensure a sufficient level of acoustic channel separation (see figure 6); otherwise, spatial arrangement of multiple speakers as specified by Auro 9.1 would be hardly useful.

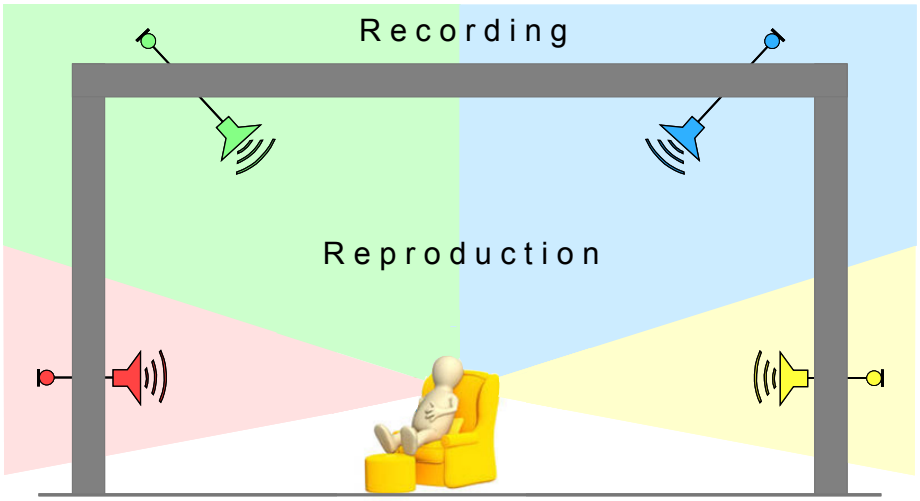


Figure 6: Reproducing the original incident directions requires strict channel separation during the recording process

There is no doubt that realizing acoustic channel separation for room miking becomes more difficult, the larger the number of playback channels is. There is an increasing risk of undesired crosstalk, i.e. correlated contents on three or more speakers. This again results in clear coloration that also depends on the listener's position within the listening environment. Placing the microphones in a way that no interfering crosstalk occurs is very difficult with nine channels! There are two solutions that also work with 5.1: Either the use of optimized techniques such as OCT surround or increase of distances between microphones and thus the propagation delays in order to alleviate crosstalk (largely spaced A/B configurations).

Figure 7 shows two sample arrangements and the highly different reflection patterns they produce. The example simulates a source (with first-order image sources) reproducing a Dirac impulse in a rectangular parallelepiped room (a "shoebox"). The figure shows the first 50 ms of the resulting signal at the sweetspot of an Auro 9.1 speaker arrangement.

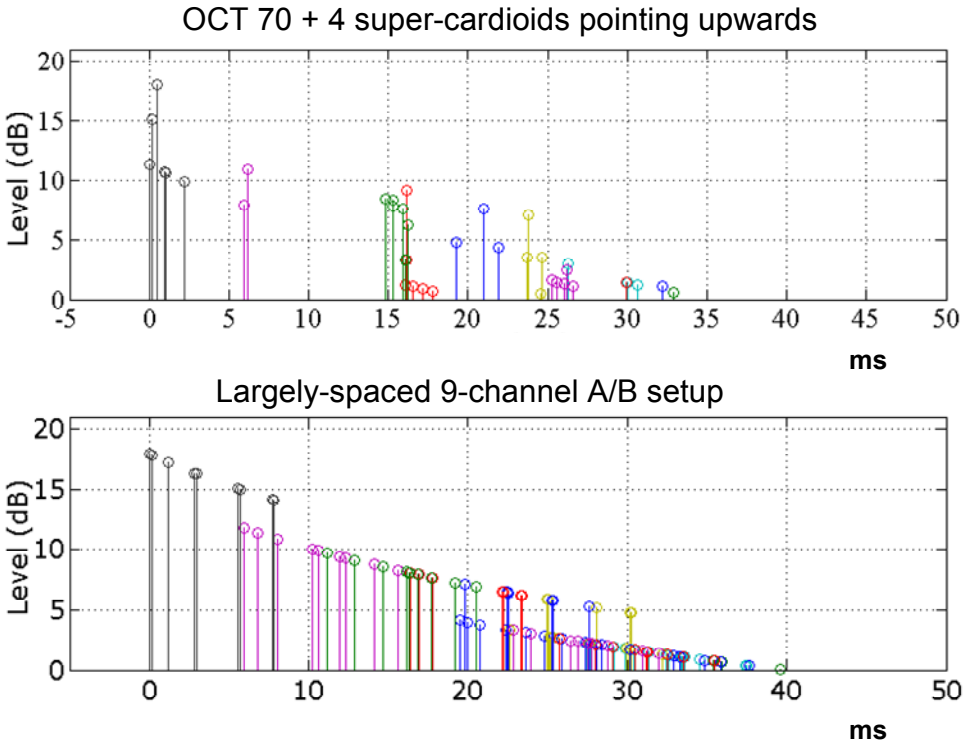


Figure 7: Reflection patterns in the sweetspot of an Auro 9.1 speaker array, generated using 2 different microphone arrays. The microphone arrays record the same source. A shoebox-shaped recording room was produced for emulation purposes. The source produces a Dirac impulse. Each peak color corresponds to a (1st order) image source.

The upper image contains the reflection patterns generated by a 9-channel arrangement similar to OCT (OCT70 plus four supercardioid microphones pointing upwards, see figure 10). Direct sound (black peaks) and the reflections produced in the recording room are reproduced with highest clarity and without any crosstalk from the direction that is consistent with the recording room. The second image shows a largely-spaced 9-channel A/B setup. The conditions are entirely different: Obviously, there are hardly any utilizable discrete reflections, each of the 9 channels contains the whole branch of early reflections from all directions ("nine times mono").

The reverb builds up very quickly, even the direct signal has a wide and reverberant character; however, this may actually be desired: Recording in long-reverb spaces where the diffuse-field (the envelopment) dominates the listening experience – for example, in a church – results in a great surround sound; presence and imaging stability can still be enhanced using spot microphones. Achieving a degree of imaging, depth, and distance perception corresponding to the recording room will definitely not be achieved.

5.2. Using Artificial or Convolution Reverb

Modern technologies would also allow for alternative approaches based on convolution. The necessary spatial information is gained either by sampling the physical recording room or existing rooms of high acoustic quality, or by using calculated models. Basically, the concept uses convolution algorithms for several locations in the area of the sources to be imaged (e.g. a stage). This allows for convolving signals from separate microphones or microphone groups with the room's IRs from specific room directions. For Auro 9.1, this requires eight convolutions per source signal (with the IRs from the eight corners of the room). Figure 8 shows the principle for a specific stage area (microphone group A).

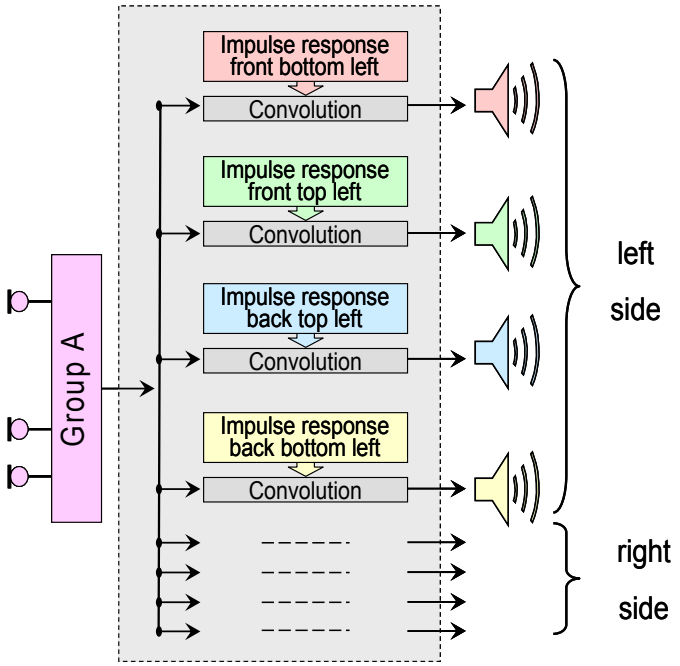


Figure 8: Concept of a convolution processor producing 8-channel early reflections

If one decides not to use model-based IRs in order to ensure realistic imaging, the IRs need to be sampled in advance using suitable directive microphones. In addition, if the microphone's directivity is not adequate, unwanted sound-incidence directions likely to cause crosstalk can be shaded. (This might also include direct sound.) Afterwards, any future recordings made in that room can be convolved with the sampled impulse IRs. If desired, the engineer might do the mix without using convolution reverb and record the diffuse-field (including background noise) using room microphones. This allows for creating a realistic balance between the reverb and applause / audience noise. The use of convolution, however, eliminates a number of practical recording problems and also provides more freedom of creativity.

5.3. Diffuse Sound

Diffuse sound (i.e. reverb or background noise) needs to be reproduced diffusely. This can be achieved using Auro 9.1 if appropriate signals are fed to the extra speakers. Diffuse signals must be sufficiently different on each speaker, that is, they need to be de-correlated over the entire frequency range. A sufficient degree of independence is necessary, in particular, in the low-frequency range as it is the basis of envelopment perception (for an example, see [14]). However, increasing the number of channels that need to be independent makes recording more complex. It is a tough job to generate de-correlated signals using first-order microphones – for example, a coincident array such as a double MS array or a Soundfield microphone allows for generating a maximum of four channels providing a sufficient degree of independence [15]. Therefore, the microphone array needs to be enlarged to ensure de-correlation.

It is worth noting here that measuring diffuse-field correlation is not trivial. There are two reasons for this: First, measuring the correlation requires the diffuse sound level to be much higher than direct and reflection levels, so the distance from the source needs to be sufficiently large. Secondly, considering the degree of correlation is not sufficient; this does not account for the fact that low-frequency (de-)correlation is particularly important [13].

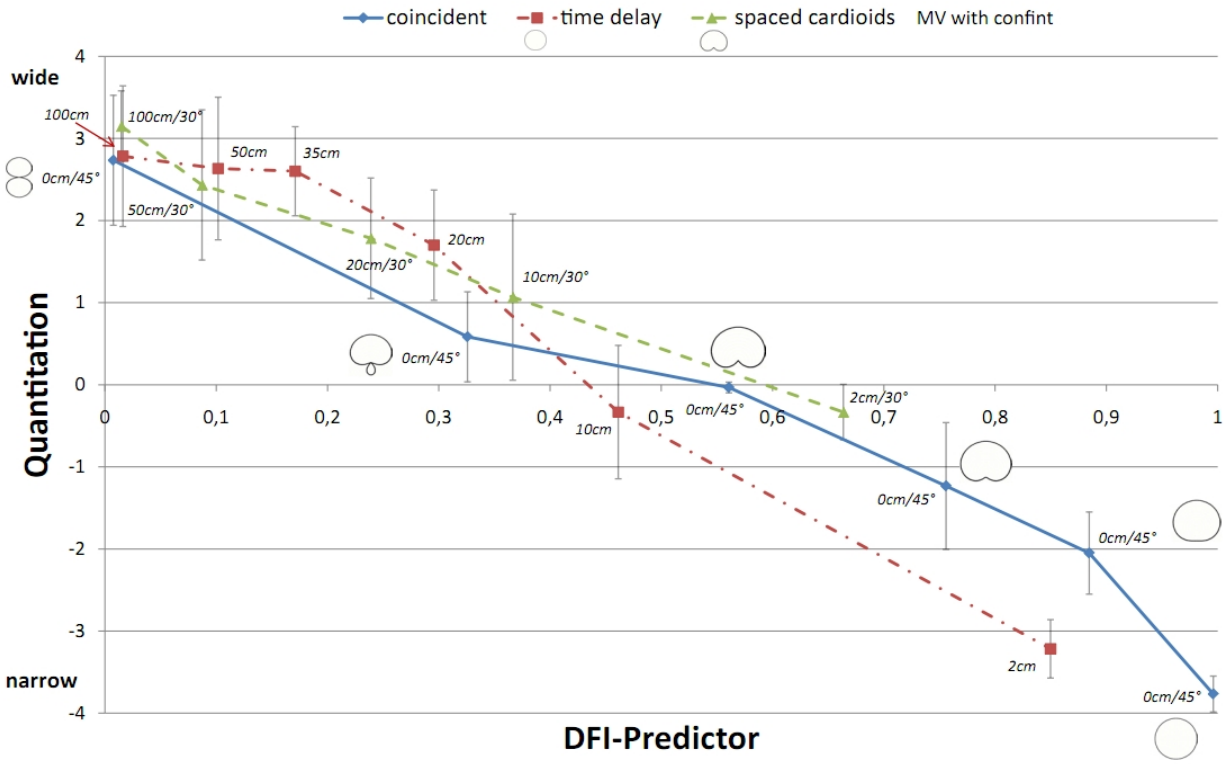


Figure 9: Interrelation of the DFI predictor and the perception of spatial width (taken from [13]). Arrays include (from left to right):
Coincidental (blue): r=0 cm, ±45°; omnidirectional portion: 0 (Blumlein), 0.4, 0.5 (cardioid, X/Y), 0.6, 0.7, 1 (mono)
Equivalent (green): Cardioid array, ±30°; spacing: 1m, 50cm, 20cm, 10cm, 2cm;
Delay (red): Omnidirectional array; spacing: 1m, 50cm, 35cm, 20cm, 10cm, 2cm

A study on the effects of diffuse-field correlation may be useful for determining the required minimum spacing and angles of microphone pairs. Coincident, equivalent, and delay-based techniques might be suitable for eliminating diffuse-field correlation (see [13]). Figure 9 shows the interrelation between the DFI predictor (a frequency-weighted degree of coherence) and the subjectively perceived stereo width. Mono portions in the diffuse-field often distract listeners due to its narrowness and the coloration they produce. Several coincident, equivalent, and delay arrays were simulated. We assume that only those arrays causing low diffuse-field correlation will be acceptable as they do not restrict the perception of spatial width (i.e. quantitation > 2). There are six arrays meeting this requirement: the Blumlein pair array (2 coincident figure-eight microphones, ±45°), two equivalent cardioid arrays, and three omnidirectional arrays (microphones spaced more than 35 cm).

6. Design of 3D Microphone Configurations for Auro 9.1

6.1. Main Microphones

The realisation and even the application of a multichannel main microphone for Auro 9.1 reproduction appear to be particularly difficult. The microphone setup has to pick up the direct sound, early reflections, reverberation, and enveloping sources (e.g. applause), and to deliver the complete nine-channel mix which must satisfy with respect to many parameters, such as sound colour, directional imaging, spatial imaging, and envelopment as described above. The parameters are governed by psychoacoustic principles and practical constraints leaving not much room to get everything well in any application. Suitable recording conditions must be given. Thus, the scope of applications for a specific main microphone is limited. Different kinds of main microphones fit different applications. For example, a main microphone designed according to a realistic reproduction as also described in this paper, fits only applications in which a natural, realistic reproduction is aimed at, e.g. the recording of a small music ensemble. The opposite can be the case for e.g. film music which may need to be unnaturally wide and not concrete in localization.

As outlined above, suitable miking techniques meeting all requirements requires sufficient prevention of interfering crosstalk. This becomes problematic with an increasing number of channels. For example, maximum separation of an eight-channel microphone configuration is only 10 dB. This can be realized with an “8 channel supercardioid hedgehog”, where each capsule is pointing to one corner of a cuboid and thus generating 12 stereophonic sections (including 4 vertical sections) among 8 loudspeakers of the Auro 9.1 setup (see also Fig. 14a). In the microphone configuration 8 channels are limiting the maximum axis opening angle to 90° (resulting channel separations of arbitrary microphone pairs can be calculated by means of the “Image Assistant” [17]). Sufficient prevention of interfering crosstalk is hardly possible with more than 8 channels.

However, we can make use of the fact that stereophonic phantom source imaging in the vertical sections cannot practically be accomplished due to instability of localization (see Chapter 4.1). Instead, filling the areas in height with conveniently de-correlated signals is possible and an important creative element. Spatially distributed individual acoustic sources (e.g. early reflections) recorded with a largely spaced A/B setup can produce many different inter-channel delays, resulting in robustness against movements in the listening area. Furthermore, regarding recording / reproduction of diffuse sound such as reverberation,

largely spaced A/B ensures maximum de-correlation even at the low frequencies and thus supports natural immersive perception of the surrounding ambience (see Chapter 5.3).

This offers the opportunity to apply well-proved 5.0 microphone setups for the ITU 5.1 surround layer ensuring optimum imaging quality such as OCT surround (see e.g. [1]), completed by an adequate microphone configuration for the height loudspeakers. Latter should be designed with a view to the given recording situation and aesthetical intention. For example, in the case of natural music recording in a concert hall it is proposed to pick up only the indirect sound from the ceiling by using 4 supercardioid microphones pointing upwards, located approximately 1 m above the OCT Surround setup and arranged in a 1 m spaced A/B square (figure 10).

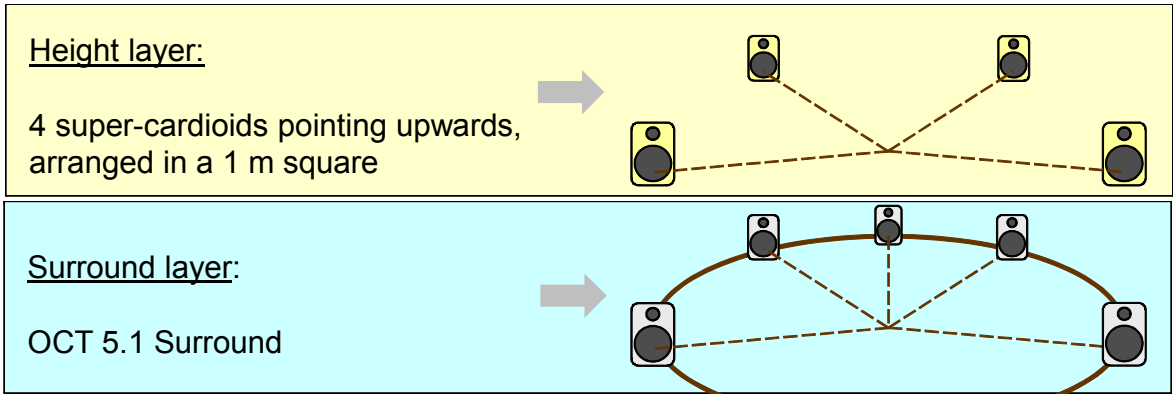


Figure 10: Example of a main microphone configuration for the Auro 9.1 format

This height arrangement fades out the direct sound from stage and lateral and rear indirect sound and noise (applause) of the auditorium on the ground as well. The resulting height sound contains both reflections from the ceiling and the corresponding diffuse portion (compare with figure 6) which is de-correlated in the frequency range above about 200 Hz.

6.2. Room and Ambience Microphones

The application of a main microphone appears to be advantageous if suitable recording conditions are given and the correct microphone location can be found to ensure the adequate directional image as well as the adequate balance of direct and indirect sound. This is even more true if naturalness and a “being there” impression is intended by means of 3D recordings. In this case, spot microphones are added in order to realize corrections with respect to the loudness balance, sound colour, etc., e.g. as described in the previous chapter, combined with delayed spot mikes according to the rules of the Room Related Balancing concept [1].

However, the spot microphones can also create a complete mix of the sources (on the stage) which meets all the requirements with respect to the directional and loudness balance. Now, a 3D room microphone is required. This should not pick up the directional sound, since directional imaging is done with the spot mix. Therefore directional stability of the stereophonic front image across the L-C-R channels is needless, and the center channel appears to be dispensable here. An Auro 9.1 room microphone configuration should feed 8 channels and meet all the requirements described above with the exception of directional

imaging attributes. A number of solutions are considerable, for example two Hamasaki Squares [1], [18], one for the lower Auro 9.1 surround layer, one for the height layer. It should be mentioned that there are approaches which generate additional 4 channels from the usual surround layer channels (e. g. upmix by introducing delay). However, experiences with 3D room sound recording are just growing.

A special situation is given with public address live recording where the early reflections are not of interest. For example, with the recording of live pop music it appears to be more important to catch the reaction of the audience and the acoustic ambience whereat the high PA sound level coming from the sides of the stage should be suppressed as much as possible. In these cases a directive ambience microphone is desirable. A cardioid configuration as shown in figure 11 is proposed. All cardioids are looking backwards in order to avoid direct sound from the stage.

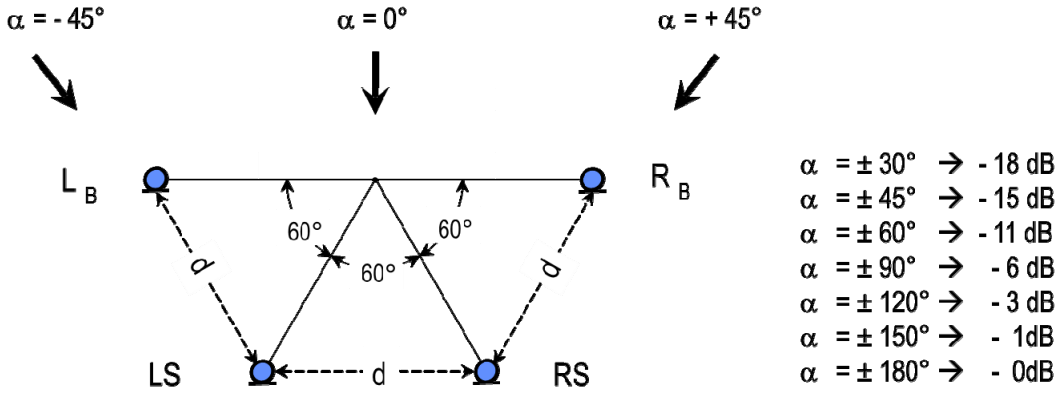


Figure 11: Cardioid trapezoid ("Theile-Trapez") for suppression of direct sound (d = 60 cm)

The suppression of the sound from the stage within the sector ± 45° is more than 15 dB and more effective compared to the above mentioned Hamasaki Square. The level of a sound source, which travels from back to front, decreases continuously according to the directivity of the cardioids. All three stereo microphone pairs L_B-LS, LS-RS and RS-R_B are working thereby in an identical manner as pure narrow-spaced A/B (ΔL = 0 dB) and are producing complete de-correlation above 300 Hz in the diffuse field. From the distances d = 60 cm results a recording angle of 60° for each of the three recording sectors (check with "Image Assistant" [17]).

Regarding 3D ambience recording the consideration of the recording situation is advisable. For example, in most of the open air live events neither any ambience sound nor elevated direct sound could usefully be reproduced from the height loudspeakers. In these cases the height microphone layer should be dispensable. In cases of in-hall PA events additional height information could support the immersive perception to a certain degree.

If a 3D ambience recording is desired, the choice could be the same solution as described in Chapter 5.4: Four supercardioids pointing upwards are mounted about 1 m above the cardioids pointing backwards; they contribute the (indirect) sound from the ceiling. An additional cardioid trapezoid feeding the height loudspeakers as shown in figure 12 will be

more effective regarding the suppression of sound from the stage. Since both microphones layers are identical they are working as an A/B configuration also in the vertical axis.

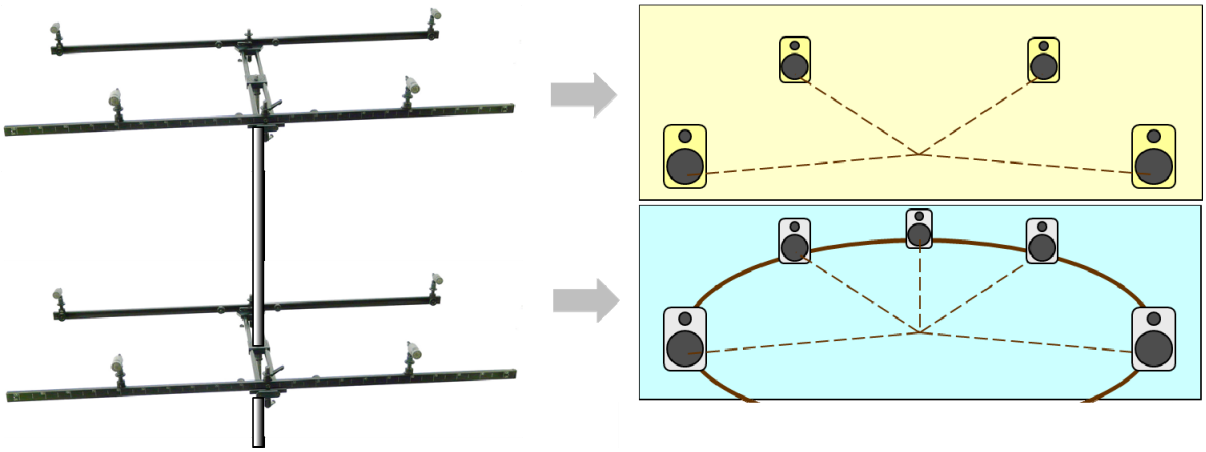


Figure 12: 3D directional ambience configuration for the Auro 9.1 format

For an ambience microphone there are often specific requirements. In [19] three ambience levels are depicted. These three levels are shown in figure 13: the diffuse sound (level 1), the discrete, but location-independent sound (level 2) and the discrete, location-dependent sound (level 3).

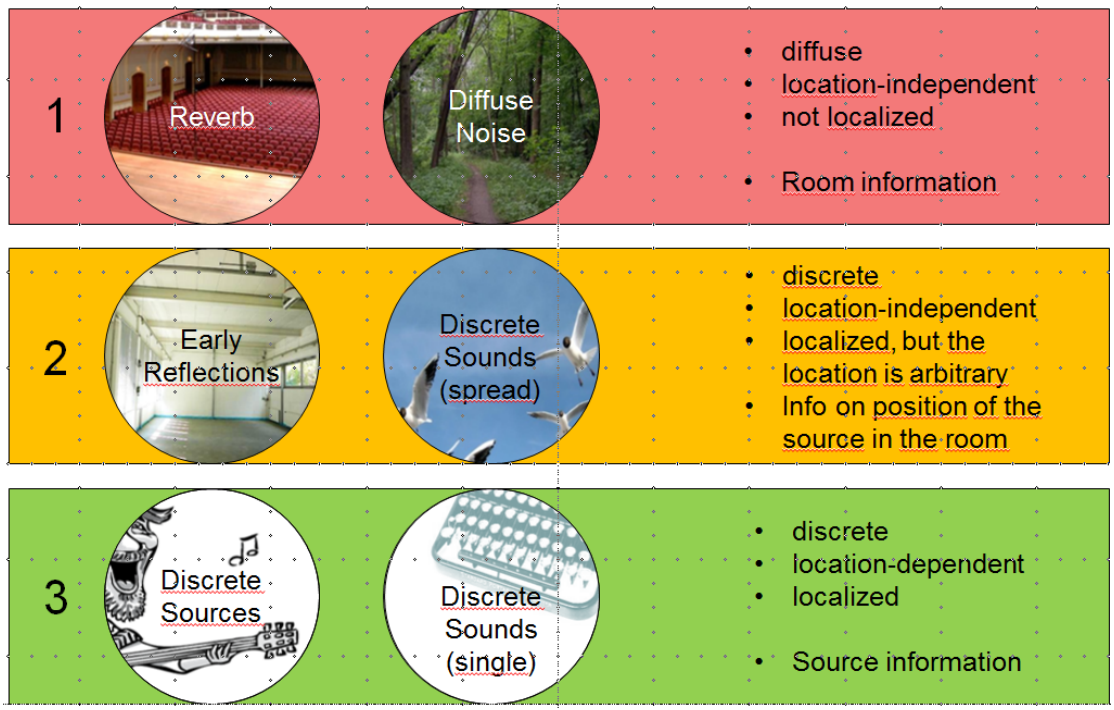
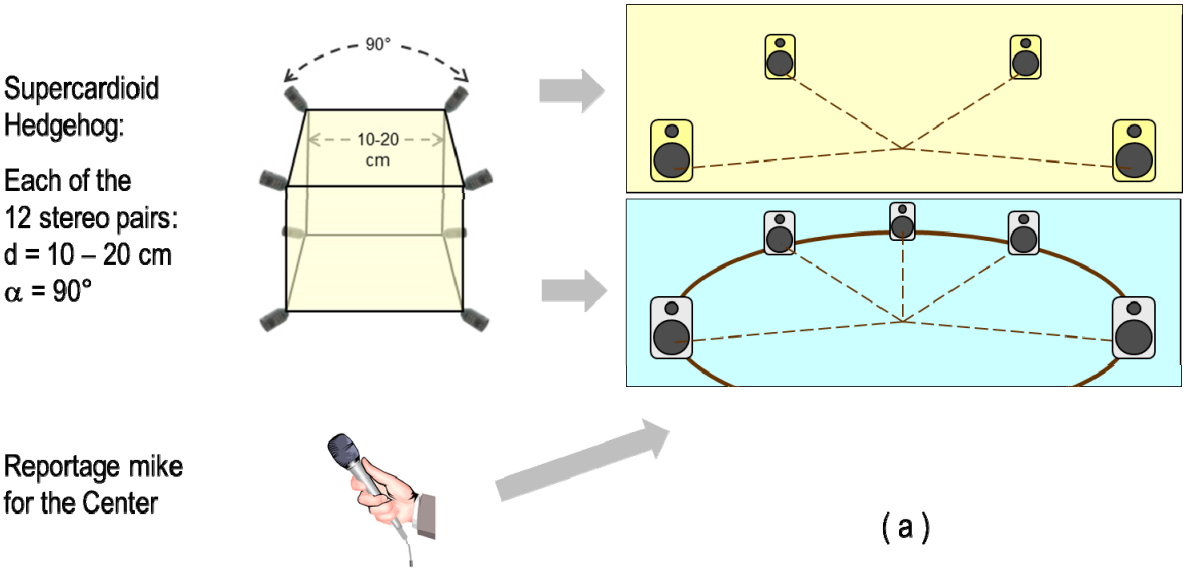
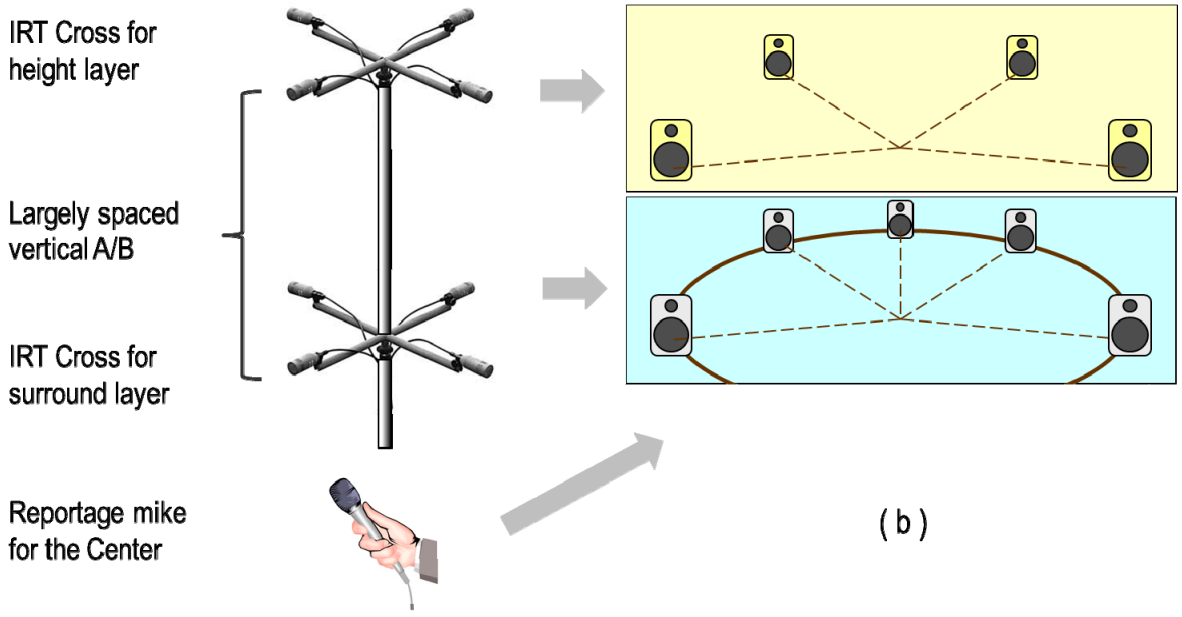


Figure 13: The 3 ambience levels, from [19]

For each ambience level and for each combination of levels in an acoustical scene, different recording techniques may be optimal. For example, a purely diffuse sound field (level 1) can be picked up by a Hamasaki square-like setup with large distances, optimized for the lowest inter-channel correlation. Whenever level 2 and/or level 3 signals are present, the directional imaging between the loudspeaker signals gets important and thus a correspondingly optimized setup may be advantageous. Examples are the "8 channel supercardioid hedgehog", as described in chapter 6.1, using an edge-length (inter-microphone spacing of 10 – 20 cm), or a 3D atmo configuration using two IRT crosses [1] as indicated in Fig. 14a and 14b.





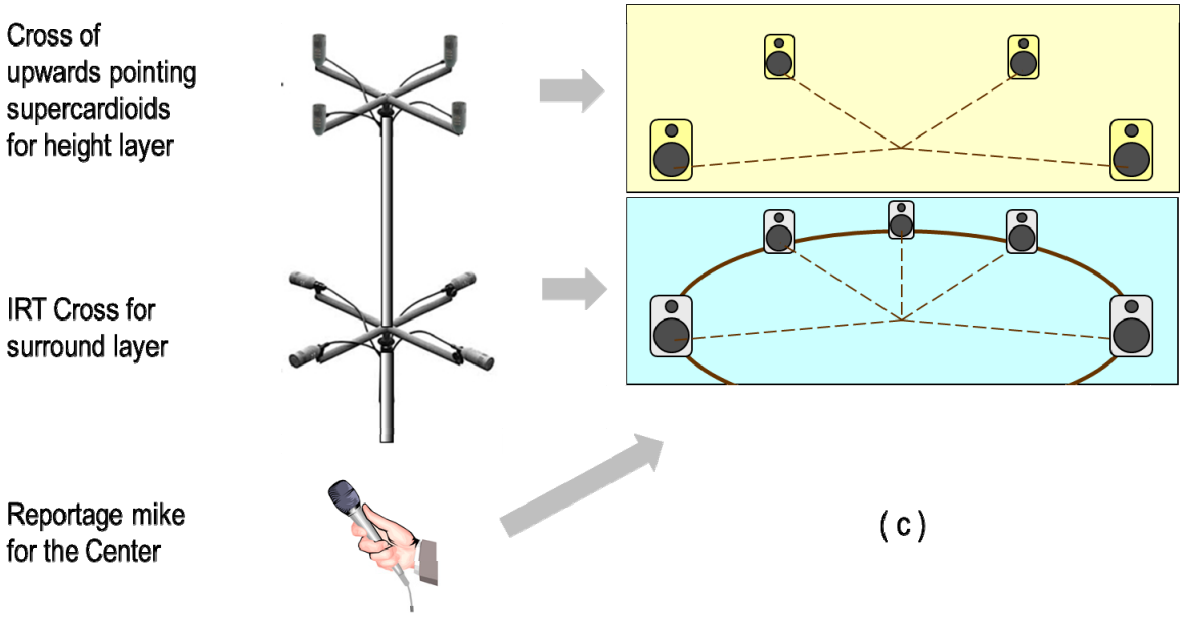


Figure 14: Three examples of 3D atmo microphone configurations for the Auro 9.1 format

The configurations can be modified in order to match requirements and intentions in individual recording situations. If imaging of discrete sources (ambience levels 2 and 3) in the height loudspeaker layer is not required the setup may change to an IRT cross consisting of four cardioids with an edge length of 25 cm for the lower layer and four upwards-pointing supercardioids in the upper layer, mounted about 1 m above the lower layer (Fig. 14c). In any case the consideration of ambient levels appears to be useful for both loudspeaker layers.

The authors would welcome very much further practical feedback and experience exchange!

7. References

[1] Theile, G.: "Natural 5.1 Music Recording Based on Psychoacoustic Principles". *Nordic Sound Symposium XX, BOLKESJØ, 2001.* *www.hauptmikrofon.de/theile/Multich_Recording_30.Oct.2001_.PDF*

[2] Wikipedia: "7.1 surround sound". *http://en.wikipedia.org/wiki/7.1_surround_sound*

[3] Dabringhaus, W.: "2+2+2 Aufnahme-Verfahren". *www.mdg.de/frame2.htm*

[4] Holman, T.: "10.2 channel surround sound". *http://en.wikipedia.org/wiki/10.2*

[5] Van Baelen, W.: "Challenges for Spatial Audio Formats in the near Future". *Tagungs-bericht 26. Tonmeistertagung 2010. (ISBN 978-3-9812830-1-3), pp. 196-205*

[6] Theile, G.: "Zur Kompatibilität von Kunstkopfsignalen mit intensitätsstereofonen Signalen bei Lautsprecherwiedergabe: Die Klangfarbe". *Rundfunktechn. Mitt. 4/1981, pp. 146-154*

[7] Steickart, H.: "Kopfbezogene Stereofonie – neuere Erfahrungen bei Produktion und Rezeption". *Tagungsbericht 15. Tonmeistertagung 1988, pp. 316-331*

[8] Dickreiter, M., Dittel, V., Hoeg, W., Wöhr, M.: *Handbuch der Tonstudiotechnik, vol. 1, chapter 5. K. G. Saur Verlag Munich, 2008 (ISBN 978-3-598-11765-7)*

[9] Wittek, H.: "Räumliche Wahrnehmung von virtuellen Quellen bei Wellenfeldsynthese". *Tagungsbericht 23. Tonmeistertagung 2004, pp. 268-297.* *http://hauptmikrofon.de/HW/Wittek_TMT2004_Paper_final.pdf*

[10] Horbach, U., Pellegrini, R., Felderhoff, U., Theile, G.: "Ein virtueller Surround Sound Abhörraum im Ü-Wagen". *Tagungsbericht 20. Tonmeistertagung 1988, pp. 238-245*

[11] IRT: "Binaural Room Synthesis BRS". *www.irt.de/de/produkte/produktion/binaural-room-synthesis-brs.html*

[12] Fey, F.: "Ohrenbetörend. Smyth Research SVS Realiser A-8". *Studio Magazin 12/2009", pp. 24-34*

[13] Riekehof-Böhmer, H., Wittek, H., Mores, R.: "Voraussage der wahrgenommenen räumlichen Breite einer beliebigen stereofonen Mikrofonanordnung", *Tagungsbericht 26. Tonmeistertagung 2010. (ISBN 978-3-9812830-1-3), pp. 481-492*

[14] Griesinger, D.: "General overview of spatial impression, envelopment, localization, and externalization". *Proceedings of the 15th International AES Conference, Copenhagen, 1998, pp. 136-149.*

[15] Wittek, H., Haut, C., Keinath, D.: "Doppel-MS – eine Surround-Aufnahmetechnik unter der Lupe". *Tagungsbericht 24. Tonmeistertagung 2006., pp. 700-731*

[16] Barbour, J.: "Elevation Perception: Phantom Images in the Vertical Hemi-sphere". *Proceedings of the 24th AES Conference on Multichannel Audio", The New Reality, June 2003*

[17] Wittek, H.: "Image Assistant". *JAVA applet available on www.hauptmikrofon.de*

[18] Hamasaki, K.; Fukada, A.; Kamekawa, T., Umeda, Y.: "A concept of multichannel sound production at NHK." *Tagungsbericht 21. Tonmeistertagung 2000, D13*

[19] Wittek, H.: " Mikrofontechniken für Atmoaufnahme in 2.0 und 5.1 und deren Eigenschaften". *Tagungsbericht 27. Tonmeistertagung 2012 (ISBN 978-3-9812830-3-7)*