

# GO BACK TO BED - THE LIMITS OF COHERENT OBJECT BASED SIGNALS IN DOLBY ATMOS

Master Thesis in Sound Design  
Daniel Eaton, +41 76 595 21 65  
Mentorship - Manu Gerber

Master of Arts in Sound Design  
Zurich University of Arts - ZHdK  
August 2023

**Z**

—  
—  
—  
—

**hdk**

Zürcher Hochschule der Künste  
Zurich University of the Arts

# ABSTRACT

In recent years, Dolby Atmos has become a mainstream technology no longer exclusive to movie theaters. The endorsement by large corporations such as Apple, Amazon, Disney+, Netflix and Tidal has made the technology a widely available consumer format for film and music content. There are however differences between consumer and cinematic Dolby Atmos setups that have an impact on playback.

The production of an immersive concert video with the ensemble Nik Bärtsch's Mobile is used as an example to show potential challenges when working with coherent multichannel signals in different setups. This thesis explores recording and mixing approaches to facilitate the accurate playback of object bed-based audio on various systems and suggests workflow adaptations to enhance compatibility when working in this format.

<b>1   IMMERSIVE AUDIO TECHNOLOGIES</b>	<b>5</b>
1 1 5.1 AND 7.1	5
1 2 AURO 3D	6
1 3 AMBISONICS	8
<b>2   DOLBY ATMOS</b>	<b>9</b>
2 1 BEDS	9
2 2 OBJECTS	10
2 3 SPATIAL CODING AND CLUSTERING	11
2 4 CINEMATIC CONTENT	12
2 4 1 SPEAKER ARRAYS	12
2 4 2 SPEAKER PLACEMENT	13
2 4 3 THE ACADEMY CURVE	13
2 5 CONSUMER CONTENT	14
2 6 OBJECTS AND DELAY	15
2 7 DIALNORM AND LOUDNESS	16
2 8 DATA FORMATS AND BINAURAL PLAYBACK	17
2 8 1 DD+ JOC	17
2 8 2 APPLE SPATIAL AUDIO	17
2 8 3 DOLBY TRUE HD	18
2 8 4 AC4-IMS	18
2 9 BINAURAL RENDERING	19
<b>3   PRODUCTION APPROACHES IN DOLBY ATMOS</b>	<b>20</b>
3 1 SPATIALIZATION, CORRELATION AND COHERENCE	20
3 2 DECOHERENT SIGNALS	21
3 3 CORRELATED SIGNALS	24
3 3 1 $\Delta T$	24
3 3 2 $\Delta L$	24
3 4 SPOT MICROPHONES	26
3 5 POSTPRODUCTION WORKFLOWS IN DOLBY ATMOS	26
3 6 SOUND AND VIDEO	29

<b>4   NIK BÄRTSCH'S MOBILE IN DOLBY ATMOS</b>	<b>30</b>
4 1 PREMISE	30
4 2 ARRAY CHOICE	31
4 2 1 2L-CUBE	31
4 2 2 KS1 SETUP	32
4 3 THE RECORDING SESSION	35
4 4 MIXING APPROACH	35
4 4 1 SPOT SOURCE CHOICES	36
4 5 OBJECT PANNING	37
4 6 MIXING THE MODULS	37
4 6 1 MODUL 65	37
4 6 2 MODUL 68	38
4 6 3 MODUL 58	38
4 6 4 MODUL 61	39
4 7 CHALLENGES	40
4 7 1 CROSSTALK	40
4 7 2 DELAY	41
<b>5   CONCLUSIONS AND OUTLOOK</b>	<b>42</b>
5 1 DECORRELATE YOUR MICROPHONE ARRAYS	42
5 2 THE LIMITATIONS OF OBJECTS	44
5 3 OUTLOOK	46
<b>6   ACKNOWLEDGEMENTS</b>	<b>47</b>
<b>7   BIBLIOGRAPHY</b>	<b>48</b>
<b>8   APPENDIX</b>	<b>51</b>
<b>9   STATEMENT OF ACADEMIC HONESTY</b>	<b>53</b>



# 1 | IMMERSIVE AUDIO TECHNOLOGIES

In this thesis the term immersive audio refers to the extension of sound sources into the third dimension. For practical purposes this usually consists of a half-dome arrangement of sources above the listener, although spherical arrangements of sound sources are possible with sound field technologies such as Ambisonics (see 1|3 *Ambisonics*).

The most common immersive stereophonic arrangements are based on 5.1 and 7.1 surround setups as these have been the industry standard for cinematic productions for decades. They have well known properties and are widely available, which facilitates the spread of the newer technologies and makes retrofits to existing setups easier. The choice of technology for immersive audio boils down to mainly three options which are each different in approach, availability and cost. This chapter provides a brief summary of the most common technologies apart from Dolby Atmos.

## 1 | 1 5.1 AND 7.1

5.1 and 7.1 setups have been the standard specified by the ITU for home (ITU, 2023) and cinema surround (Dolby, 2015) for decades. The addition of a center speaker [C] to a stereo [L] and [R] setup opens up the possibility of having a sound source anchored to the screen. Stereo setups are capable of producing an accurate stereo field if the listening position is along the equidistant axis between the speakers but produce phase differences and a skewed image once the listening position is off center. Physically centering mono signals while freeing up the left, right, and the surround channels for other audio content allows for increased panning precision and an extension of the sweet spot (Dickreiter, Dittel, Hoeg, & Wöhr, S. 229). The presence of surround speakers [Ls] [Rs] [Lrs] [Rrs] furthermore expands the available room information to the sides and rear. In a stereo setup, the simulated room information is constrained to the area between [L] and [R] (Dickreiter, Dittel, Hoeg, & Wöhr, S. 242). The added [LFE] channel is generally used for low frequency effects and is not used as a subwoofer for the full-range 5.0 setup.

Surround channels in cinema theatres are attenuated by 3dB with respect to the front channels [L], [C] and [R] (Dolby, 2011, p. 2).

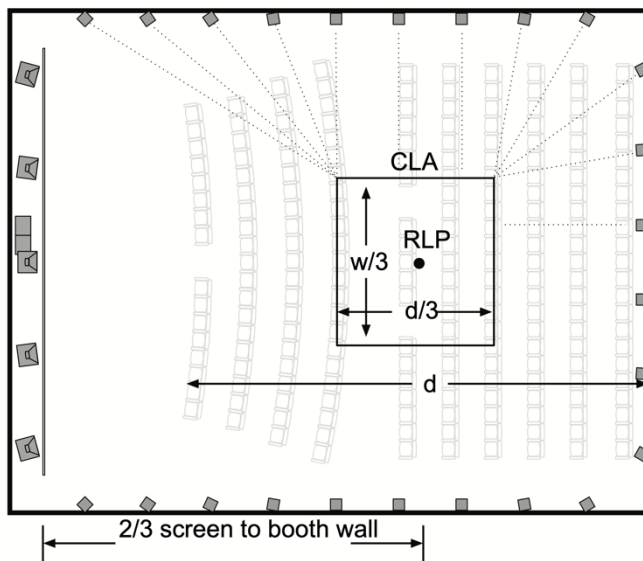
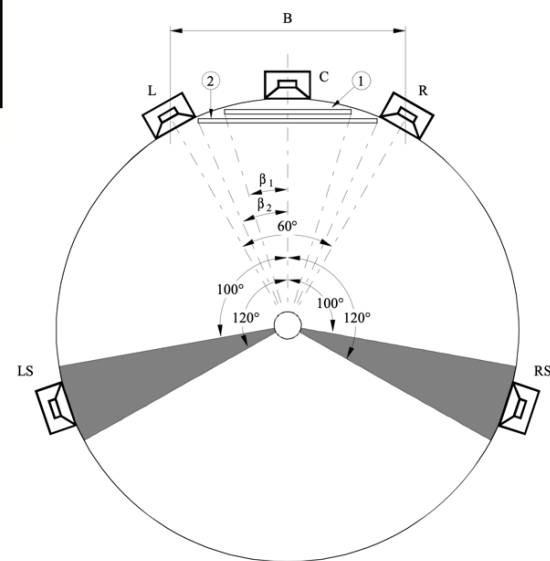


Figure 1 An example for Dolby's recommended cinema surround speaker array setup. The central listening area (CLA) around the reference listening point (RLP) is shown in relation to the auditorium width  $w$  and the distance  $d$  between the first and last row (Dolby, 2015, p. 15)

Figure 2 The recommended ITU-R BS.775-1 reference loudspeaker arrangement for a [L][C][R][Ls][Rs] setup.  $B$  denotes the loudspeaker base width and  $H$  the height of the screen. The gray area shows the tolerance for the angle of the surround speakers. (ITU, 2023, p. 3)



Screen 1 HDTV – Reference distance =  $3H$  ( $2\beta_1 = 33^\circ$ )  
Screen 2 =  $2H$  ( $2\beta_2 = 48^\circ$ )

## 1 | 2 AURO 3D

Auro 3D was developed by the Belgian company Auro Technologies in order to expand the 5.1 and 7.1 surround formats into the third dimension by adding a height layer and a top layer (Voice of God) to create a sonic half dome. Similarly to Dolby Atmos, Auro 3D uses a combination of beds and objects to deliver scalable immersive audio content. The speaker placement of the height layer in Auro 3D is at an angle between  $25^\circ$  and  $35^\circ$ . While our brains are capable of creating accurate phantom images in the horizontal plane up to an angle of approximately  $70^\circ$  between speakers, they are only able to do so up to an angle of about  $40^\circ$  in the vertical plane. At larger angles we lose the perception of natural coherence between the signals and perceive them as two individual sound

sources. The reason for this seems to be our inability to perceive time-of-arrival differences in the vertical plane (Pfanzagl-Cardone, 2023, pp. 101-102).

Auro 3D encodes the height channels into the surround channels, which can be played back as a downmixed surround format. These nearly lossless PCM streams contain all the height information and can be decoded by a suitable playback system without the need of separate downmixes. One advantage is an absence of audio artifacts due to lossy compression like the ones in Dolby Atmos consumer formats. A further advantage is the relatively compact data size which makes lossless immersive live streams possible (Pfanzagl-Cardone, 2023, pp. 131-134).

Auro 3D was developed as a consumer as well as a cinema format, which entails different requirements when it comes to room setups and equalization. Consumer playback systems are ideally set up in a way that all speakers are equidistant from the listening position while having a flat frequency (Pfanzagl-Cardone, 2023, p. 104). In contrast, cinema mixes should take into account the «X-Curve» or «Academy Curve» of the speakers' frequency response as well as the array-based playback (see 2|4|3 *The Academy Curve*).

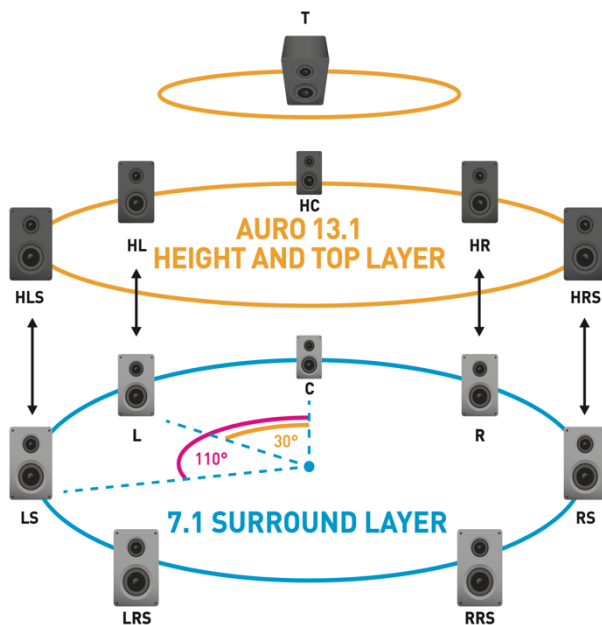
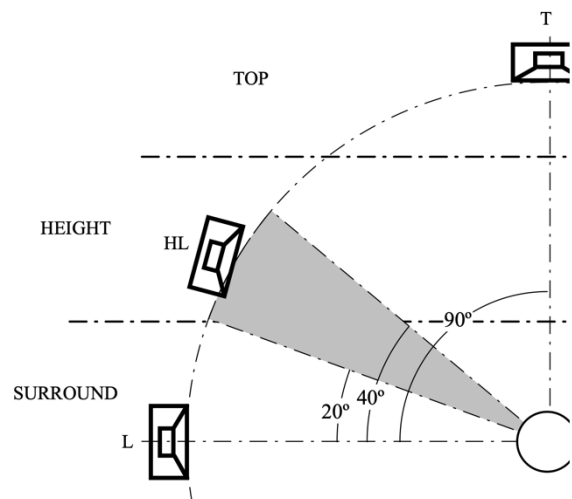


Figure 4 Height and top layer angles in an Auro 3D setup. The range of the vertical angle should be between 20° and 40°. (Auro Technologies, 2023, p. 19)

Figure 3 Example of a 13.1 Auro 3D setup. The height and top layer channels are encoded in the lower channels to reduce bandwidth without perceivably compromising audio quality. (Auro Technologies, 2023, p. 14)



## 1 | 3 AMBISONICS

In contrast to the stereophonic approaches discussed in this thesis, Ambisonics is a sound field technology. This means that the individual speakers emit partial signals representing the desired sound field when combined at the listening position. This technology aims to recreate the sound field at one specific point, whereby the resolution is given by the so called «Ambisonics order» denoting the order of the spherical harmonic component of the sound field.

This technology has the advantage of having a spherical symmetry as well as relying less on specific speaker setups like the other technologies in this thesis. While excelling at reproduction of coincident higher order Ambisonics (HOA) recordings, spaced microphone array recording methods cannot be reproduced without significant comb filtering, as the individual speakers of the playback setup emit highly correlated signals with time differences. HOA are excellent for placement of point sources in an isotropic 3D space and are therefore ideal for VR content played back over headphones and electroacoustic composition (Pfanzagl-Cardone, 2023, pp. 192-195). The recordings for this thesis were made with spaced microphone arrays and therefore carried room information. This made this particular technology unsuitable for the production.

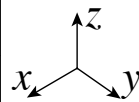

















































$l$ :		$P_\ell^m(\cos \theta) \cos(m\varphi)$							$P_\ell^{ m }(\cos \theta) \sin( m \varphi)$						
0	$s$														
1	$p$	 													
2	$d$	  							 						
3	$f$	   							  						
4	$g$	    							   						
5	$h$	     							    						
6	$i$	      							     						
$m$ :		6	5	4	3	2	1	0	-1	-2	-3	-4	-5	-6	

Figure 5 An overview of the real spherical harmonics of the order  $[l]$  and degree  $[m]$ . Similarly to the Fourier expansion of a signal, Ambisonics describes a sound field through its spherical components (Wikipedia, 2023).

## 2 | DOLBY ATMOS

The technology of Dolby Atmos was chosen for this thesis for its widespread availability as an immersive audio format for cinematic and consumer content.

Dolby originally developed Atmos for cinema by expanding the 7.1 channel-based format into the third dimension and introducing the concept of beds and objects (Pfanzagl-Cardone, 2023, pp. 143-146). One important difference to the 5.1 and 7.1 Dolby Digital format is the absence of the -3dB attenuation of the surround channels in order to facilitate panning between the front channels and the rest of the speaker setup.

### 2 | 1 BEDS

Beds represent the traditional channel-based formats from 2.0 up to 7.1.2. These beds can be used in the same manner as before, as they are routed to the corresponding speakers in the listening environment or adapt to speaker setups according to the downmix settings as specified in the Dolby Atmos Renderer. Panning in beds will produce phantom images in the same way as working in two-dimensional stereophonic surround formats would.

Beds are channel-based premixes or stems that include multichannel panning via phantom imaging, and do not need dedicated panning via Dolby Atmos metadata. Beds are bound to fixed locations in space tightly constrained to traditional speaker environments, including theatrical environments where generally speaker arrays are used for surround channels (Dolby, 2023).

Beds can be used for stereo-adjacent workflows such as bus processing and their behavior is well understood. Speaker placement in channel-based workflows has been specified for optimal playback with the respective channel counts (Gray, 2023).

All beds in a master file share the same binaural setting, while objects can all have different settings. The individual bed channels however can have different settings (Dolby, 2023).

## 2 | 2 OBJECTS

Objects in Dolby Atmos are mono (stereo linkable) audio sources which can be positioned at ear level and above and carry channel adaptive Object Audio Metadata (OAMD) (Gray, 2023).

Objects are discrete audio elements that can be placed anywhere in the three-dimensional sound field. They can be used to position audio content more precisely than with bed panning. Objects can utilize as few, or as many, speakers as defined by their positional and size metadata and can be static or moving (Dolby, 2023).

They remain separate tracks in an Atmos master file and get rendered by the playback system, making the format scalable. They are the only way to create separation of front and rear height channels, as Dolby's maximum bed size limits the top layer to left/right panning. Dolby Atmos does not treat bed and object streams differently when they are panned to discrete speakers. For sources outside of these positions however, an object-based workflow is preferential, as the Object Audio Renderer's (OAR) spatial rendering capabilities supersede the phantom image generated in beds (Gray, 2023).

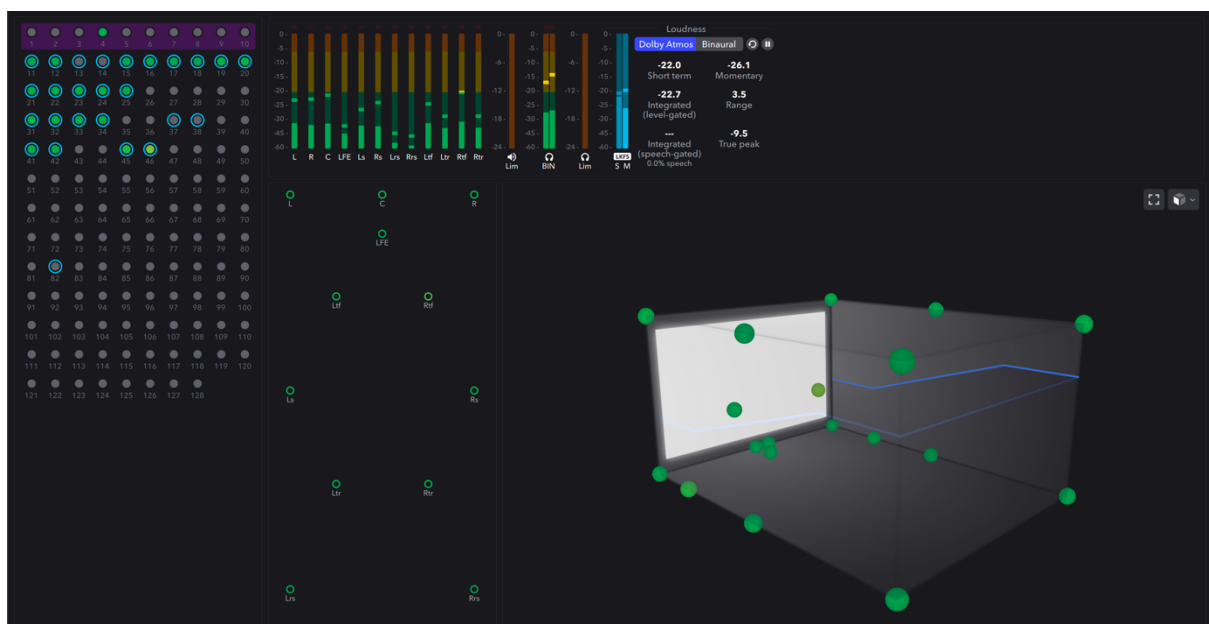
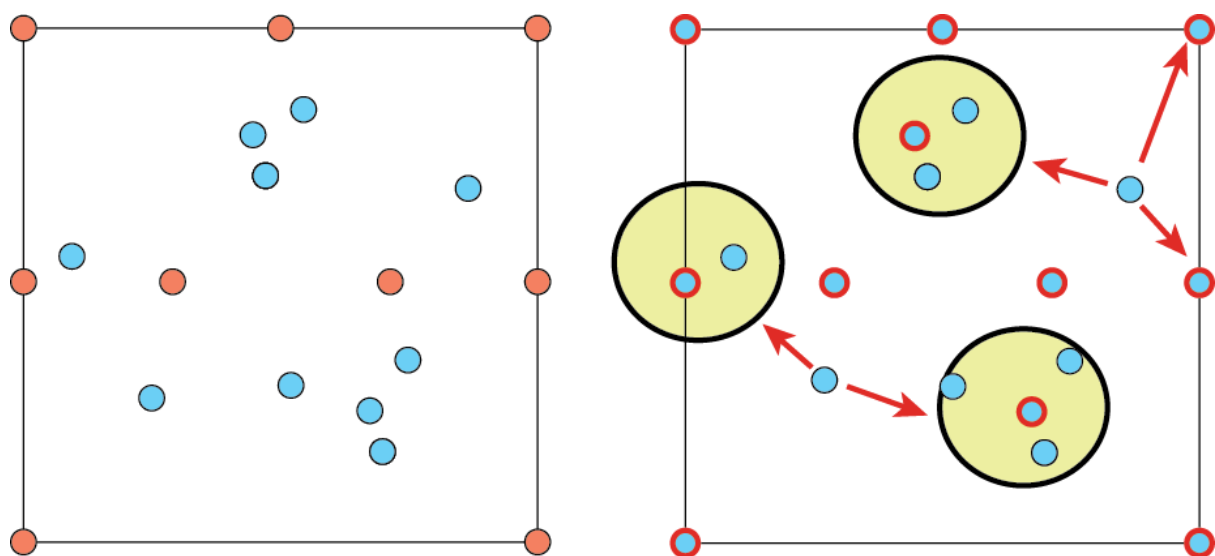


Figure 6 Screenshot of the Dolby Atmos Renderer showing an overview of the active channels in a master file. The only active channel in the bed (purple) is the LFE. The blue rings around the object channels symbolize the reception of panning data while the dots' colors show the audio levels from green to red.

## 2|3 SPATIAL CODING AND CLUSTERING

For consumer formats, every channel apart from the LFE in a Dolby Atmos master file gets rendered as an object with corresponding object audio metadata. This happens independently of it of being part of a bed or being an object. Dolby Atmos uses a maximum of 16 spatial elements for streaming and playback to reduce bandwidth. This means that once the number of channels in a master file exceeds 16 (15 channels + LFE), the file is subject to clustering (Dolby, 2023).

The spatial coding process takes as an input a full Dolby Atmos mix (one 7.1.2 bed plus a maximum of 118 objects), and outputs a configurable number (12, 14 or 16) of output objects. This is achieved by dynamically grouping nearby objects into spatial clusters. These spatial clusters are aggregate objects, which combine the audio signal of the original objects to one position (Dolby, 2023).



*Figure 7 On the left, the original presentation (the Dolby Atmos mix without spatial coding) includes nine bed channels (in red) and ten objects (in blue). The spatial coding (right) dynamically and optimally aggregates the beds and objects into a target number of clusters (here, 11 clusters with representative position highlighted in red). Some clusters can comprise several original objects, or be the combination of an original bed and one or more original objects. Some original objects can also be redistributed among multiple clusters (Dolby, 2023).*

This process is updated once every frame, which means 24 times per second for audio content, as specified by current music delivery requirements (24fps -18LUFS integrated -1dBTP) (Gray, 2023).

The loss of quality is masked by psychoacoustic effects and can be easily perceived when soloing speaker channels during playback. Spatial coding is applied to the DD+ JOC and in a slightly different manner to Dolby TrueHD formats but is not applied to the AC4-IMS stream (see 2|8|4 AC4-IMS). It is recommended to switch off spatial coding emulation during production as not to mask panning details. This feature can be activated at a later stage of the production in order to gauge what the mix will sound like on consumer playback systems (Gray, 2023).

## 2|4 CINEMATIC CONTENT

For cinematic content, the most common method is to have several beds acting as the traditional dialogue (DX), effects (FX), and music (MX) stems. These usually contain the main channel-based content (e.g., speech in the [C] channel plus the corresponding surround reverbs and delays) with objects used for precise dynamic placement of sound sources. Producing cinematic content via the dedicated Dolby Atmos RMU takes into account the speaker arrays present in a movie theatre as opposed to discrete speakers in smaller setups. Individual speakers can be accessed via objects, but beds are played back via the speaker arrays to provide accurate channel-based playback in the venue. The cinema renderer has more computational power than the home renderer implementation and applies size and decorrelation filters during the decoding process (Dolby, 2023). This leads to some differences when comparing the workflow to consumer formats.

### 2|4|1 SPEAKER ARRAYS

Speaker arrays in cinemas consist of multiple speakers playing back the signal of a single channel. This effectively smears out the localization of the channel in question and softens the transients of the signal (see 3|3|1  $\Delta t$ ). The benefit of widening the listening area however outweighs the downside of losing precision and is thus standard practice in movie theatres with clear specifications (Dolby, 2015, pp. 4-5). As mentioned above, the cinema renderer applies a so-called size/decorrelation prebaking for optimal playback over arrays (Dolby, 2023).



## 2|4|2 SPEAKER PLACEMENT

Dolby Atmos being an extension of the existing 7.1 cinema format means that speaker placement in movie theatres is above ear level. This room geometry dependent placement of speakers automatically leads to a feeling of sound emanating from above, which is different compared to consumer playback setups (see 2|5 Consumer Content). This can lead to differences in height panning perception, which might require adaptations in the mix.

## 2|4|3 THE ACADEMY CURVE

As previously mentioned in the chapter about Auro 3D, movie theatres are not required to have a linear frequency response but mostly adhere to the so-called Academy Curve or X-Curve as specified in ISO 2969:1987/SMPTE ST 202:2010.

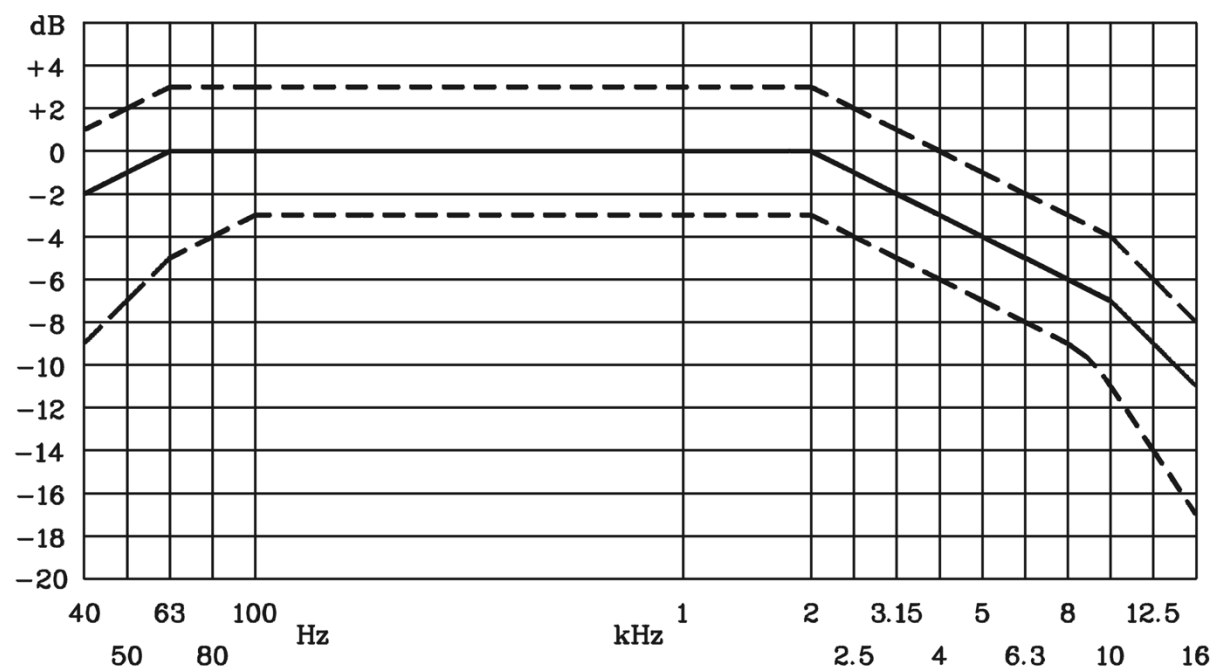


Figure 8 The academy curve shows the specified frequency response of the playback system. The dotted lines show the tolerance of the curve which allows for a margin of error of  $\pm 3$  dB between 100 Hz and 2 kHz. The attenuation of 3 dB/octave for frequencies above 2 kHz is clearly visible (Allen, 2006)

Therefore, the sound of the master mixed for cinema is not necessarily linear and does not directly translate to consumer formats. Additionally, the dynamic range used for the cinema mix is established with 85 dB (A) reference level for the front channels at the reference listening place (at 2/3 of the distance between the screen speakers and rear

wall loudspeakers (see Figure 1) of a -20 dB pink noise signal (Pfanzagl-Cardone, 2023, p. 125).

The use of the X-Curve is somewhat controversial. At its inception, it was intended to reduce the hiss of the optical sound-on-film track. Playback systems have dramatically changed and improved since then and have made this curve to tackle the hiss somewhat obsolete. However, the industry tends to adhere to this practice in order to mitigate room resonances in the bass frequencies and to compensate for a perceived harshness of a flat frequency response. There are efforts to re-evaluate the need for the X-Curve, but as of today, there is no new standard for cinemas to adhere to (Majidimehr, 2023).

## 2|5 CONSUMER CONTENT

Dolby's recommendations for playback of Atmos consumer content is based on the ITU Standard for speaker placement (ITU, 2023), meaning that ideally, all speakers (except the LFE) are placed at equal distances to the listening position as specified in the following images:

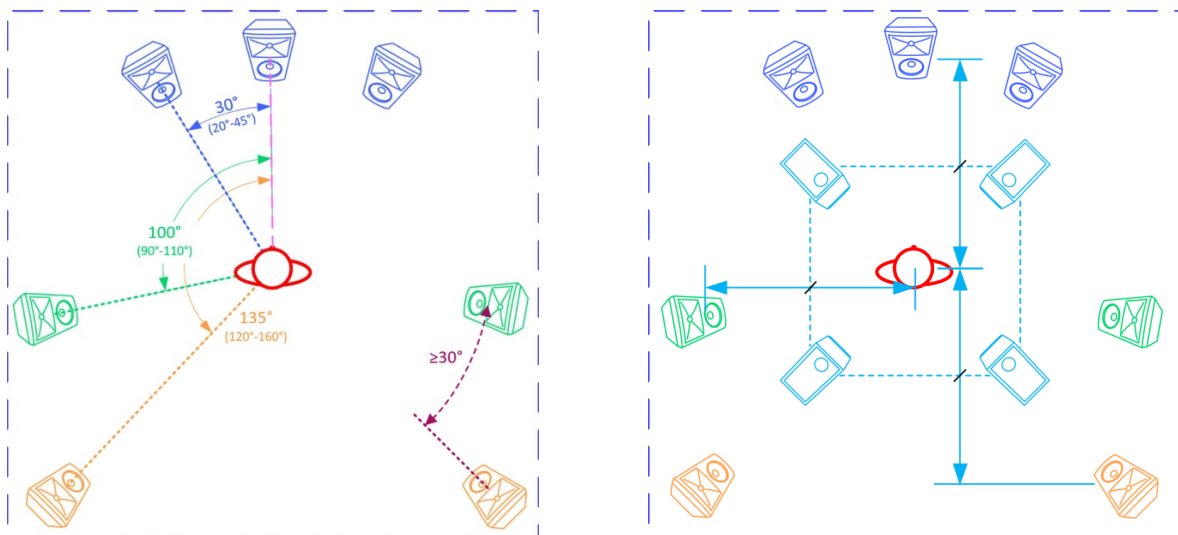


Figure 9 Suggested 7.0.4 speaker setup for a home playback system based on the ITU recommendation (Dolby, 2023)

In addition to this well-controlled setup assuring equal timing of sound sources, the ITU specification calls for a linear frequency response. This and the addition of headphone

playback as the probably most common consumer playback form require separate workflows with some additional steps and considerations.

## 2 | 6 OBJECTS AND DELAY

An important point to note is the absence of time alignment for objects in Dolby Atmos. Beds in cinema master files get played back via calibrated speaker arrays that mitigate timing differences in order to maximize the size of the sweet spot. Objects, according to Dolby, are treated differently:

«The Dolby Atmos processor in the theater intelligently assigns each audio track. It maps the bed channels to screen channels or surround arrays, and positions objects within the room. It's all reproduced in real time based on where the loudspeakers are. Dolby Atmos scales to the specific speaker complement of a theater, so the effects will be the same regardless of the auditorium's size. Sound placement is consistent throughout the audience. Thanks to audio objects originating from specific locations rather than general areas, you'll hear the exact same effect no matter where you sit in the theater – every seat is the "sweet spot."... »

«...also, in traditional surround setups, a sound moved from the screen to the surround zones drops in volume. Dolby Atmos, using improved room equalization and better bass management along with the independently powered speakers, avoids this problem. Sounds maintain the right volume as they move, adding to the realism.» (Dolby, 2023)

This poses a significant challenge for object-based mixes (see 5 | *Conclusions and Outlook*), as cross-correlated audio signals from discrete speakers do not arrive at the sweet spot simultaneously as they would in a consumer playback environment (Dolby's «every seat is the sweet spot» claim is of course physically impossible). This leads to audibly delayed transient signals and phase issues which are highly dependent on the listening position and room geometry.

## 2|7 DIALNORM AND LOUDNESS

Dialnorm (dialogue normalization) is a parameter included in all Dolby Atmos bitstreams and carries the loudness measurement. Dialnorm contains information about true peak measurements plus the integrated loudness (LUFS) of a Dolby Atmos file. This measurement is taken from a 5.1 re-render of the file in question and is the norm set by Dolby for loudness measurements. The reason for this is Atmos being a scalable format and thus requiring some standardized method of measurement for consumer formats to make normalization possible. The dialnorm value depends on the downmix and trim settings inside the Dolby Atmos Renderer, which means that the measured loudness of an Atmos mix can be changed without there being any audible difference when monitored in channel counts above 5.1. The dialnorm value cannot be defined manually as it was the case for the AC-3 codec but is written into the metadata by the Dolby Atmos Renderer. Dialnorm is dialogue gated for content with dialogue, but in the case of musical content is calculated by the Atmos Renderer through the integrated loudness measurement (Gray, 2023).

Stereo/Binaural loudness and true peak levels are also measured post limiter and are equally affected by downmix and trim settings.

The Binaural (for AC-4 IMS) and Spatial Coding (for DD+ JOC) limiters inside the renderer are a simulation of the limiting applied to an export through the Dolby Media Encoder. The soft clipping limiter is there to avoid clipping during playback. All these limiters have no effect on the Atmos master file.

It is important to note that accurate loudness measurements in Dolby Atmos can only be done offline within the Dolby Atmos Renderer. The integrated loudness in LUFS and the true peak values of an Atmos master file are equally important to normalization procedures in consumer formats (Gray, 2023).

Values above 0dBTP are possible for .mxf files for DCPs, as they correspond to the Atmos master file. They however will hit the built-in brick wall limiter of the playback system to avoid clipping (Baltensperger, 2023).

## 2|8 DATA FORMATS AND BINAURAL PLAYBACK

As previously mentioned, consumer Dolby Atmos content consists of compressed audio streams in order to reduce bandwidth. The following data formats are currently available for the consumer market:

### 2|8|1 DD+ JOC

DD+ JOC (Dolby Digital+ Joint Object Coding) is used for speaker playback and is the standard format for exporting .mp4 files from the Atmos Renderer. This format carries up to 16 audio streams at a constant bandwidth of 448 - 768kbps. It is channel adaptive, meaning it can handle up to 15 cluster positions plus the LFE that can be anything from traditional, channel-based audio to Atmos mixes. This playback format is used by Apple, Tidal, and AV receivers for speaker playback and does not carry binaural metadata (Gray, 2023). The loudness requirements for music content are -18dB LUFS integrated with -1dBTP (Dolby, 2023).

### 2|8|2 APPLE SPATIAL AUDIO

Apple requires an ADM BWF (Dolby Atmos Master) file that they use to generate a DD+ JOC file with the Dolby Media Encoder. This file then gets delivered to consumers for speaker playback. For binaural playback on headphones, they further convert this DD+ JOC file to a 7.1.4 file which is then used for their head tracked binauralization for compatible headphones. This means that non-compatible headphones currently cannot play back binauralized Atmos content via Apple services.

When the soundcheck function on Apple devices is enabled, audio gets normalized up to -16dB LUFS -1dBTP, but without an additional limiting process. This means that a -18dB LUFS -2dBTP file will get normalized to -17dB LUFS -1dBTP (Gray, 2023).

## 2 | 8 | 3 D O L B Y T R U E H D

Dolby TrueHD is a lossless format developed primarily for Blu-ray DVDs to make high resolution playback possible for the consumer market. The loudness recommendation for this format is the same as for DD+ JOC, but it has a maximum bandwidth of 18Mbps (vs. 768kbps) and a maximum channel count of 16 channels. This for example makes it possible to combine a 7.1.4 mix plus a stereo and binaural mix into a single bitstream (Dolby, 2023).

As of the completion of this thesis, Dolby TrueHD is not available outside of the Blu-ray market, but it is an elegant technology to deliver lossless immersive audio to a wider audience.

It is important to note that the distribution of ADM BWF/Atmos Master files to consumers will most likely not be an option at any point. The reason for this is the metadata included in these files, which makes distributing them the equivalent of sending multi-track master tapes to consumers (Baltensperger, 2023).

## 2 | 8 | 4 A C 4 - I M S

AC4-IMS is an immersive stereo format developed by Dolby. It is used by Tidal and Amazon Music for headphone playback and phone speaker playback. This two-channel bitstream carries binaural metadata which is used for headphone rendering, but also for the aforementioned phone speaker playback, creating an immersive experience beyond stereo (speaker virtualized). Stereo playback (Lo/Ro) is also possible (Dolby, 2021, pp. 36-37).

AC4-IMS files are created directly from the ADM BWF file and are not subject to spatial coding, giving them a more accurate spatial representation compared to the previous formats. They are very close to the binaural render of the Dolby Atmos Renderer when taking into account the loudness and normalization procedures of this file type. These files get attenuated to -16dB LUFS binaural loudness pre-encoder, which also sets the dialnorm to -16dB LUFS and are then normalized to the playback devices' loudness target. This means, that a -16dB LUFS -1dBTP file gets limited by the playback device if

the target is below that loudness value. This is a big difference to the previous codecs one should be aware of.

AC4-IMS offers platform-independent binaural playback without the need of head tracking hardware (Gray, 2023). The integration of personalized HRTFs and head tracking is not available for this format, but these features will be available in the upcoming A-JOC implementation of AC4 (Baltensperger, 2023). This could make AC4 the most widespread consumer format for Dolby Atmos content as it would cater to all playback setups with one codec.

## 2|9 BINAURAL RENDERING

The binaural metadata for headphone rendering can be set either in the Dolby Atmos Renderer or via the designated plugin inside a DAW. The metadata contains distance information that cannot be changed over the length of a master file. The distance options near, mid and far can be thought of as measures of the virtualized distance between an object or bed and the listener's head (Dolby, 2023). If the distance option is set to off, the source is not binauralized and gets played back as a regular stereo or mono signal.

Binaural mixes therefore allow the impression of a sound source being closer than in speaker setups, as there is a negligible physical distance between them and the listener. Headphone playback differs from speaker playback by completely separating the left and right channels and is therefore not stereo. This means that all interactions between sound waves taking place between the listener's ears and the physical speakers are absent, leading to a difference in perception of sources when comparing these two playback methods.

The binaural settings in Dolby Atmos make it possible to simulate source distances to the ear for the best listening experience.

The accuracy of these renders depends on the HRTF (Head Related Transfer Function) as well as the headphones used for listening (Dickreiter, Dittel, Hoeg, & Wöhr, 2014, S. 375). This is a topic beyond the scope of this thesis and will not be discussed any further.

## 3 | PRODUCTION APPROACHES IN DOLBY ATMOS

Dolby Atmos offers several technical and creative approaches to mixing. There are some important differences in productions for cinema theatres when compared to consumer content which must be taken into consideration. These differences might require separate master files for playback in movie theatres and consumer playback setups. The following technical aspects should be well understood in order to make sure that the produced content is played back accurately on the systems it was intended for.

The music industry currently adheres to the loudness standard of -18dB LUFS integrated with -1dBTP and 24fps. The streaming industry has their own loudness requirements of -27dB  $\pm$ 2dB LUFS dialogue gated, -2dBTP in a studio conforming to ITU-R BS.1770-1 for the ADM BWF (Netflix, 2023) while my research found no regulation for cinematic content.

### 3|1 SPATIALIZATION, CORRELATION AND COHERENCE

There are numerous approaches when it comes to spatialization of audio. They all depend on the method the audio was recorded or synthesized and are subject to technical constraints, and most importantly the aesthetic and sonic vision of the people creating the content.

There is a large catalog of music recordings available on streaming services which were originally produced in stereo, but have since been adapted to Dolby Atmos. As the technology has not been around for very long, especially for music content, there are no standard procedures to adhere to. Existing recordings can be upmixed with plugins such as Halo by Nugen, they can be expanded with decorrelated signals generated with immersive delays and/or reverbs, audio stems can be panned in 3D space, or any combination of these methods can be applied.

There are also many approaches to recording and producing content for immersive audio. One key point to consider is whether spatial information should be included during



the recording process or if it should be generated in post-production. This differentiation applies to all audio content, not just music, and is just as important for film.

## 3|2 DECOHERENT SIGNALS

Not including spatial information and minimizing crosstalk in the recording process yields decoherent audio material which can be freely panned in stereo or 3D space (artificial stereophony). This results in a very stable sound field which carries no intrinsic room information due to the absence of phantom imaging and comb filtering. Additionally, a decorrelated reverb or delay effect can be used to give the source material spatial context without destabilizing it, as there is still no spatial information from the recording present (Dickreiter, Dittel, Hoeg, & Wöhr, 2014, S. 294-295). The resulting room information is set by panning the audio source and applying delay and reverb effects. Stereo objects are treated as two mono objects in the renderer and possible correlation should be considered when working with them.

It is possible to record highly decoherent signals with coincident as well as spaced arrays. For this, one must consider whether it is a near field or far field recording done in the free or diffuse field. The following graph shows the frequency-dependent magnitude squared coherence function between two ideal omni microphones in an ideal diffuse field at three distances:

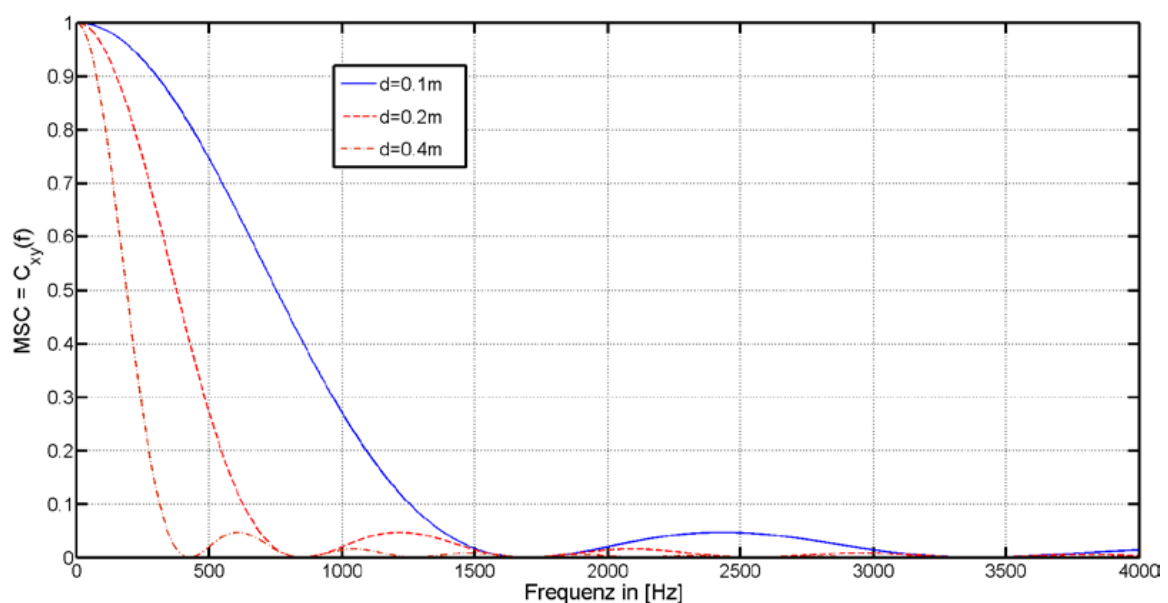


Figure 10 Magnitude squared coherence function of two omni microphones. Larger distances lead to faster frequency dependent decorrelation (Riekehof-Böhmer, 2010, S. 4).

Furthermore, the microphone directivity has a large impact on signal separation. Omnis mics will produce the most natural sound, but using them will result in more crosstalk between the individual channels. However, it is not vital for microphones to be spaced. Depending on their directivities, coincident microphones can be used to create decorrelated recordings. The following image shows the correlation between two ideal coincident microphones of the same directivity in a diffuse field. The x-axis shows the angle between the microphones while the y-axis shows the correlation between the two recorded signals. The value  $[a]$  is a description of the directivity of the capsule. ( $a=0$ ) describes a pure pressure gradient sensor while ( $a=1$ ) describes a pressure sensor.

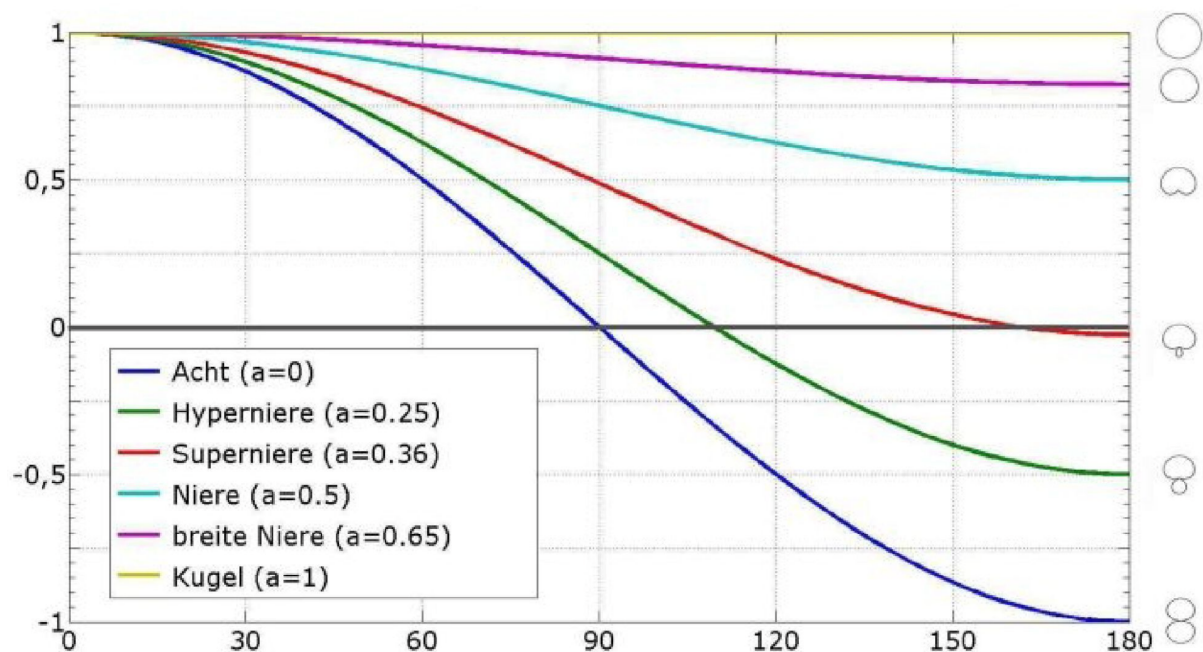


Figure 11 Correlation (y-axis) of various microphone types depending on the angle (x-axis) (Boenigk, 2010, S. 21)

It is important to understand the differences between correlation, cross-correlation, and coherence. Correlation describes the phase relation of two signals at a single point in time. The cross-correlation function takes into account the temporal differences between two signals (see 3|3|1  $\Delta t$ ), and coherence describes the maximum absolute value of the cross-correlation function (Riekehof-Böhmer, 2010, S. 2-6).

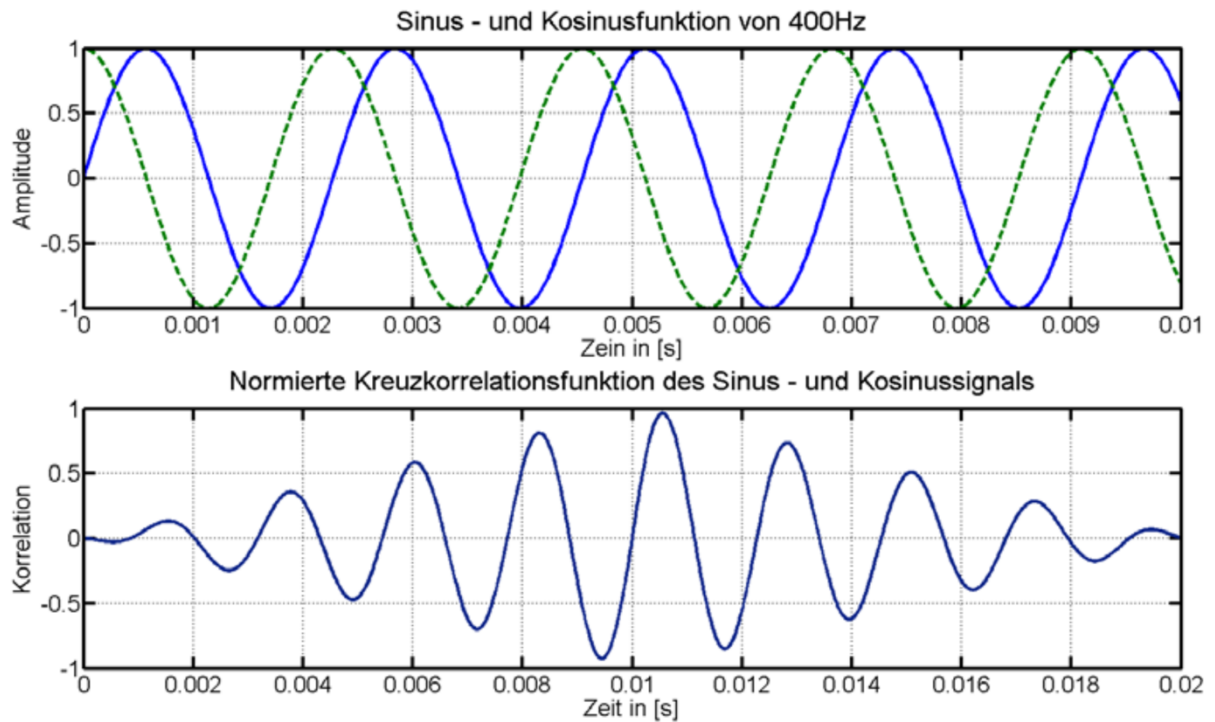


Figure 12 The upper half of this figure considers two sine waves with the same frequency and a phase difference of  $\pi/2$ . These signals are completely decorrelated. The normalized cross-correlation function of these signals is shown on the lower half of the figure. At  $t=0$  the correlation between the signals is zero, while the correlation moves between -1 and 1 when calculating the cross-correlation. This example would yield a coherence value of 1 (Riekehof-Böhmer, 2010, S. 3).

## 3 | 3 CORRELATED SIGNALS

Correlated signals contain spatial information and can be captured using the following techniques:

### 3 | 3 | 1 $\Delta T$

Spaced microphone arrays are ideal for capturing depth of field. When using spaced microphone arrays for recording, the time difference(s) of the original source signal's arrival [ $\Delta t$ ] result in a phantom image for  $\Delta t < 2\text{ms}$ . For  $\Delta t$  between 3ms and 30ms the signal is perceived as a single sound source with a growing illusion of size but with decreasing localization. This is commonly referred to as the Haas-effect or the law of the first wavefront. While these values for  $\Delta t$  are dependent on the signal spectrum, signals with  $\Delta t > 40\text{ms}$  are generally perceived as echoes (Dickreiter, Dittel, Hoeg, & Wöhr, 2014, S. 224).

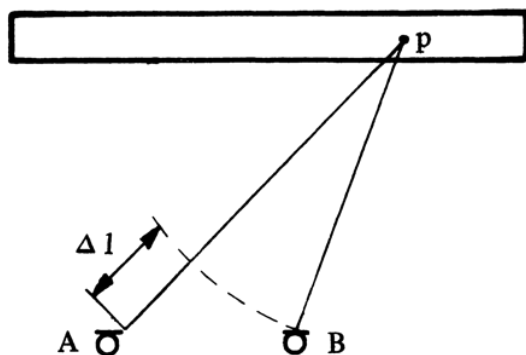


Figure 13 Sound from a source  $p$  arriving at different times in an AB array due to a distance difference  $\Delta l$  (Dickreiter, Dittel, Hoeg, & Wöhr, 2014, S. 257).

$\Delta t$  stereophony has negligible level differences between signals and therefore yields a frequency-dependent coherent signal (see Figure 10).

### 3 | 3 | 2 $\Delta L$

Phantom images in coincident microphone arrays are created purely by level differences [ $\Delta L$ ] between signals, whereby  $\Delta L > 15\text{dB}$  leads to perception of a fully panned signal in stereo (Dickreiter, Dittel, Hoeg, & Wöhr, 2014, S. 222). As an example, the side signal in a mid-side recording can be raised 15dB above the mid signal to achieve full panning of a

panned source while negative correlation must be avoided (Dickreiter, Dittel, Hoeg, & Wöhr, 2014, S. 255). This leads to very low coherence between the two channels.

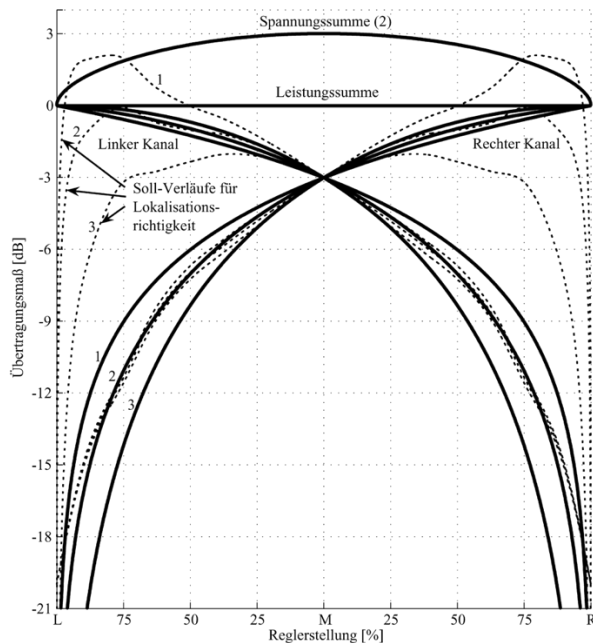
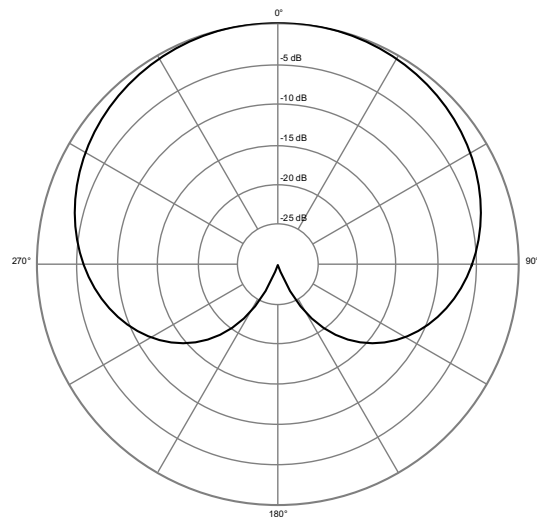


Figure 14 Panning perception diagram comparing panning settings (x-axis) vs. level differences (y-axis) between two channels. Experiments have shown that  $\Delta L > 15\text{dB}$  is sufficient for full panning of a signal (Dickreiter, Dittel, Hoeg, & Wöhr, 2014, S. 373).

Figure 15 Example of an ideal cardioid pattern. Directional attenuation can help to achieve decoherence during recording (Wikipedia, 2023)



Coincident recordings do not have frequency dependent coherence like spaced arrays, which makes them superior for precise source localization. At the same time, the depth of field is only represented by air absorption of frequencies. Omni microphones are not well suited for this technique, which leads to spectral compromises due to the directivity of the microphone capsules.

A combination of these two methods such as ORTF can be used in mixed stereophony where the benefits and drawbacks of the  $[\Delta t]$  and  $[\Delta L]$  are combined to achieve the desired results. The detailed description of these methods is beyond the scope of this thesis and will be omitted.

## 3|4 SPOT MICROPHONES

Microphone arrays are used to capture the spatial context of a sound source and produce more or less cross-correlated signals carrying room information. This means that in order to accurately represent the source signals' placement, they are usually placed at a considerable distance to the sound sources, as the diffuse field carries the room information. This makes near-field recordings impossible and leads to reduced directness and precision. To compensate for these effects, near field spot microphones are additionally set up and can greatly improve clarity of a recording when added to the mix. Although not completely decoherent due to crosstalk, they are focused on individual sound sources and can be used for artificial stereophony in real world applications. This means that the sources' panning in the array signal can be matched easily. It is important to electronically compensate for time differences between the array and spot microphone signals when combining them to maintain a homogenous sonic image (Dickreiter, Dittel, Hoeg, & Wöhr, 2014, S. 269-273).

## 3|5 POSTPRODUCTION WORKFLOWS IN DOLBY ATMOS

Dolby Atmos offers several technical approaches as well as aesthetic considerations going beyond the scope of stereo production. On the technical side there are several possibilities for using beds and objects.

- Bed-based workflow:
  - o Limited to a maximum channel layout of 7.1.2
  - o One 7.1.2 bed per Atmos file is required
  - o Panning is achieved via phantom sources
  - o Using only two top layer channels might cause issues in downmixing, as the same signal gets folded down to multiple sources in the bottom layer
  - o Limited binaural settings

- Object-based workflow:
  - o Bed is only used for LFE signals
  - o Audio is freely panned in 3D space via objects without necessarily taking into account the speaker positions
  - o Panning is achieved via the Object Audio Renderer's spatial rendering
  - o Complete control over binaural settings for every object
- Object bed-based workflow
  - o Bed is only used for LFE signals
  - o Static objects represent speaker positions and make channel layouts beyond 7.1.2 possible
  - o Better cross-compatibility (a 9.1.6 object bed can be represented on a 7.1.2 system)
  - o Additional moving objects or alternate static positions are possible
- Hybrid approach
  - o Bed can be used beyond the LFE channel for reverbs etc. to separate effects from the dry audio
  - o Bed can be used to free up object channels if needed (maximum 118 objects possible)

On the recording side, there are also some fundamentally different approaches leading to vastly different results. The following methods are equally valid for stereo and immersive formats.

- Microphone arrays with added spot microphones
  - o Accurate room information from the main array
  - o Spot microphones help with localization and sonic precision
  - o Most "natural" sound
  - o Often used in classical music productions
  - o Static sonic image
  - o Mostly strongly coherent signals

- Spot/close microphone recording
  - o Artificial stereophony necessary
  - o Minimal room information
  - o If required, room information needs to be artificially generated
  - o Close and transient-rich sound (usually near-field)
  - o Panning can change freely
  - o Mostly decoherent signals

For formats beyond stereo, the sound sources can move outside of the constraints of stereo playback. While panning in stereo outside of  $\pm 30^\circ$  is impossible, surround formats move from a simulated spatial perception to real spatial perception. Although phantom sources are unstable at these speaker angles as we saw previously, the early reflections physically emanating from the sides provide the desired effect of immersion and are less sweet spot dependent. These early reflections are vital for spatial perception in artificial room simulations and array recordings. There is not more room information in a surround format, but the room information is no longer constrained to the playback through [L] and [R] and thus  $\pm 30^\circ$  (Dickreiter, Dittel, Hoeg, & Wöhr, 2014, S. 240-242).

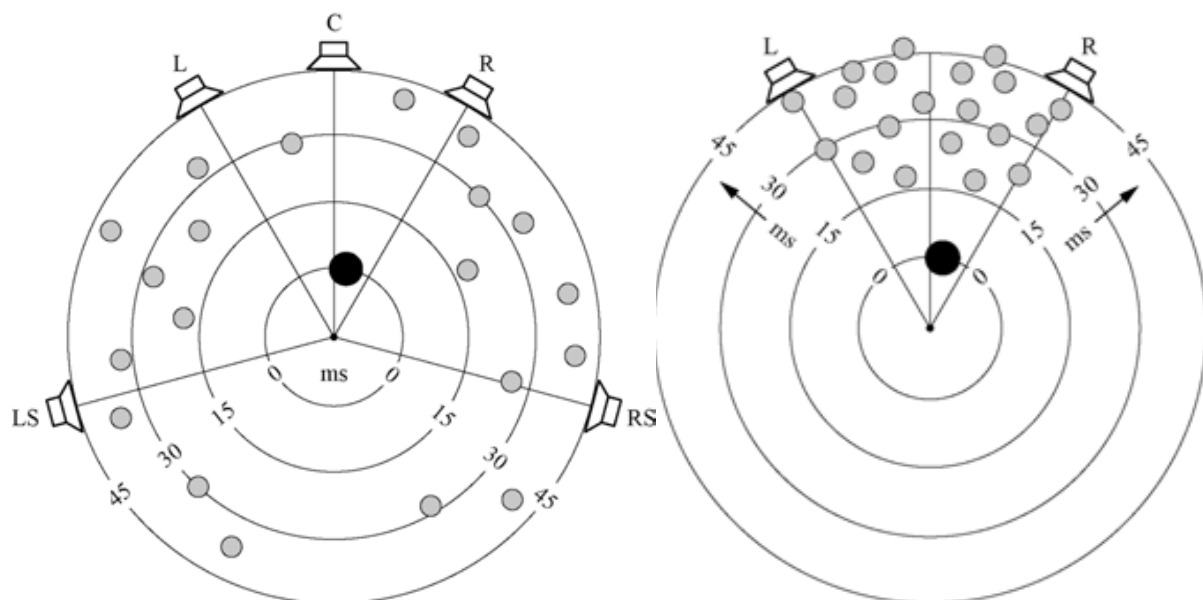


Figure 16 Spatial and temporal distribution of early reflections in two-channel and five-channel stereophony. While constrained to the angle between [L] and [R] in two-channel stereophony, multichannel stereophony opens up the full listener's plane for sound sources (Dickreiter, Dittel, Hoeg, & Wöhr, 2014, S. 240-242).



## 3|6 SOUND AND VIDEO

Video content greatly influences our perception of sound and vice versa. This cognitive processing of multimodal stimuli (Dasovich-Wilson, Thompson, & Saarikallio, 2022) is a highly complex topic beyond the scope of this thesis. However, it is a key consideration when working with video and sound and should be briefly mentioned.

While the human brain can determine the spatial origin of a sound source, it cannot give us accurate information about the space surrounding it. There are clues about the room geometry and materials contained in the reflections which makes us aware of spatial relations. Exact information about the space in which a sound propagates cannot be perceived by the human auditory system. By contrast, our visual perception lets us perceive very accurate readings of the shapes and dimensions involved and thus greatly influences our perception when added to a purely auditive experience (Di Stefano, 2022). Consider perfectly synchronized video and sound feeling unnatural once a sound source is displayed at a distance. While technically correct, there is a cognitive discrepancy because the human brain never perceives sound and light synchronously at a distance. One of the key tasks when working with sound in film is to account for these perceptual nuances while maintaining a homogenous mix. Completely adapting timing and panning to video edits produces a perceptually correct audiovisual perspective, but the constant changes in sound yield an erratic mix which is highly distracting to the viewer (Dickreiter, Dittel, Hoeg, & Wöhr, 2014, S. 926f).

It is important to note that there is most likely no one correct way of doing this. Every audiovisual content has its own characteristics that sound can help to transport and finding a good balance is in the hands of the sound designer or mixing engineer. Taste and experience play an important role in this process and cognitive dissonance can deliberately be used to create a feeling of unease if desired. Our senses usually tell us when something feels off and are probably the best indication for a skewed spatiality. As an example, Modul 68 recorded for this thesis was first mixed in stereo without the video and had to be adapted after the video edit was available, because there was a discrepancy between the roomy visual component and the rather dry and direct sound of the first mix.

## 4 | NIK BÄRTSCH'S MOBILE IN DOLBY ATMOS

The previous chapters in this thesis have been a summary of technology, theory and common practice. This was necessary to establish the background for understanding the decisions, challenges and restrictions during the work process.

### 4 | 1 P R E M I S E

Having worked with immersive audio on several projects during my studies in sound design and having a musical background, I wanted to combine these aspects of my work for my master thesis. I asked Nik Bärtsch's Mobile, an ensemble I've worked with for years, if they were interested in an immersive audiovisual production to explore the possibilities of this rather new format. The band consists of highly skilled players (Nik Bärtsch on piano, Sha on reeds and percussion, and Nicolas Stocker on drums and percussion) with a strong musical vision which made a speedy recording session possible. The music has a high dynamic range and combines strong rhythmic passages with ambient soundscapes and buildups, making it an ideal collection of sonic content for experimentation.

After considering several location options we decided to do the recording session at Konzertsaal 1 (KS1) at the Toni Areal, as it offered an adequate sonic quality for acoustic music while also having good visual properties for the video. We decided to record four previously unrecorded pieces named Modul 65, Modul 58, Modul 61 and Modul 68.

## 4|2 ARRAY CHOICE

After considering several approaches to immersive audio recording, I decided to use a variation of Morten Lindberg's 2L-Cube as the main array. This setup using omni microphones provided the most natural spectral quality and the best depth of field representation in a test comparing several immersive audio arrays in the VdT Magazin (Gericke & Mielke, 2022).

### 4|2|1 2 L - C U B E

The 2L-Cube was developed by Morten Lindberg for his channel-based immersive audio recordings. His Grammy award-winning purist approach does not allow any process applied in post-production apart from editing. His goal is to sonically balance everything prior to recording through choice of the venue, ideal placement of the main array, and the placement of musicians and instruments to achieve an optimal balance with usually no spot microphones. The array consists of a cube of variable size with one omni microphone at each corner plus an omni microphone in the middle of one of the sides for the center channel. These are then routed to individual speakers in a 5.0.4 setup. Lindberg has since expanded this setup with two additional channels for 7.0.4, whereby these are placed slightly offset and break the cube geometry (Inglis, 2022).

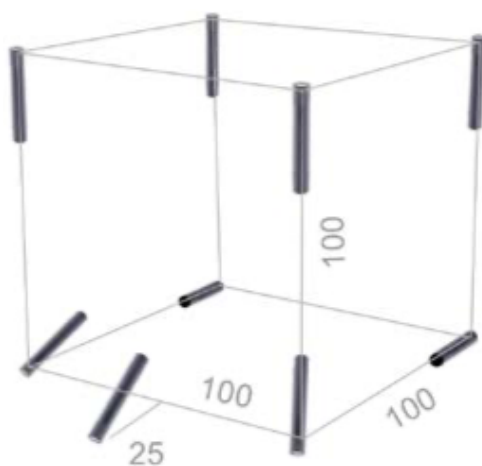


Figure 17 Rough dimensional drawing of the 2L-Cube in a 5.0.4 setup with omni microphones (Gericke & Mielke, 2022, S. 36).

## 4 | 2 | 2 K S 1 S E T U P

The modification of the 2L-Cube as the main array was to place five omni microphones on a circle (150cm diameter) in order to further approximate the ITU speaker setup when playing back the file. The angles between the microphones do not match the angles between speakers, but the distances between them (57cm and 105cm) provide an adequately blurry localization for this method to work well with spot microphones.

It is important to note that the chosen dimensions are close to the upper limit for accurate phantom imaging in the front channels. Distances above 65cm between channels mean  $\Delta t > 0.2\text{ms}$  and result in an unstable phantom image (see 3|3|1  $\Delta t$ ). This is apparent in Decca Tree recordings that usually need artificial stereophony from spot microphones to compensate for the holes in the stereo image.

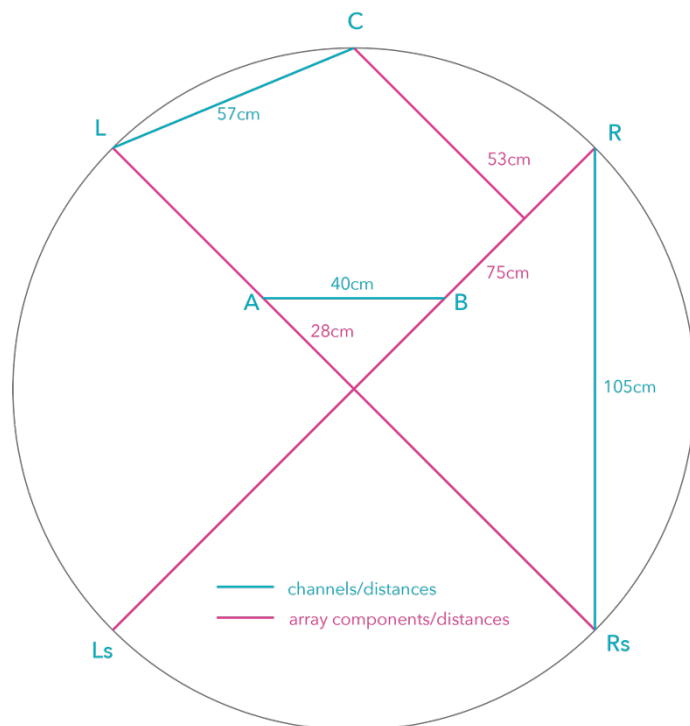


Figure 18 Dimensional drawing of the 5.0 array. The blue lines show the distances while the red lines show the array components as seen in Figure 19 (own graphic).

The array with an additional AB pair added for comparison was suspended at a height of 2.8m at a distance of 1.8m-2.5m to the musicians. The top layer consisted of four cardioid microphones (for signal isolation) pointed upwards at a height of 4.6m with an equal spacing of 1.8m. Two additional omni microphones were placed at the rear end of the room at a height of 6m spaced 3.6m to capture a very distant and diffuse signal. Near-field signals were captured with a variety of spot microphones, whereby several techniques and types were used in order to compare the different characteristics of the

microphones and their respective behavior in a Atmos mix (see 8 | *Appendix for a complete list*). This semi-open concept with an abundance of microphones was chosen not only to make comparisons possible, but to have redundancy when it comes to sources. There was a time constraint for this recording and only stereo monitoring available at the time, so this was the most realistic option to make flexibility possible in post-production. It was a deliberate choice to record the performance in a way that the musicians did not rely on personal headphone monitoring but were close together to provide optimal communication and visual compactness. This setup allowed for the best musicality and freedom, but also led to significant crosstalk between channels.



*Figure 19 View of the modified 2L-Cube with the additional AB array. One of the truss-mounted top layer microphones can be seen (own photograph).*

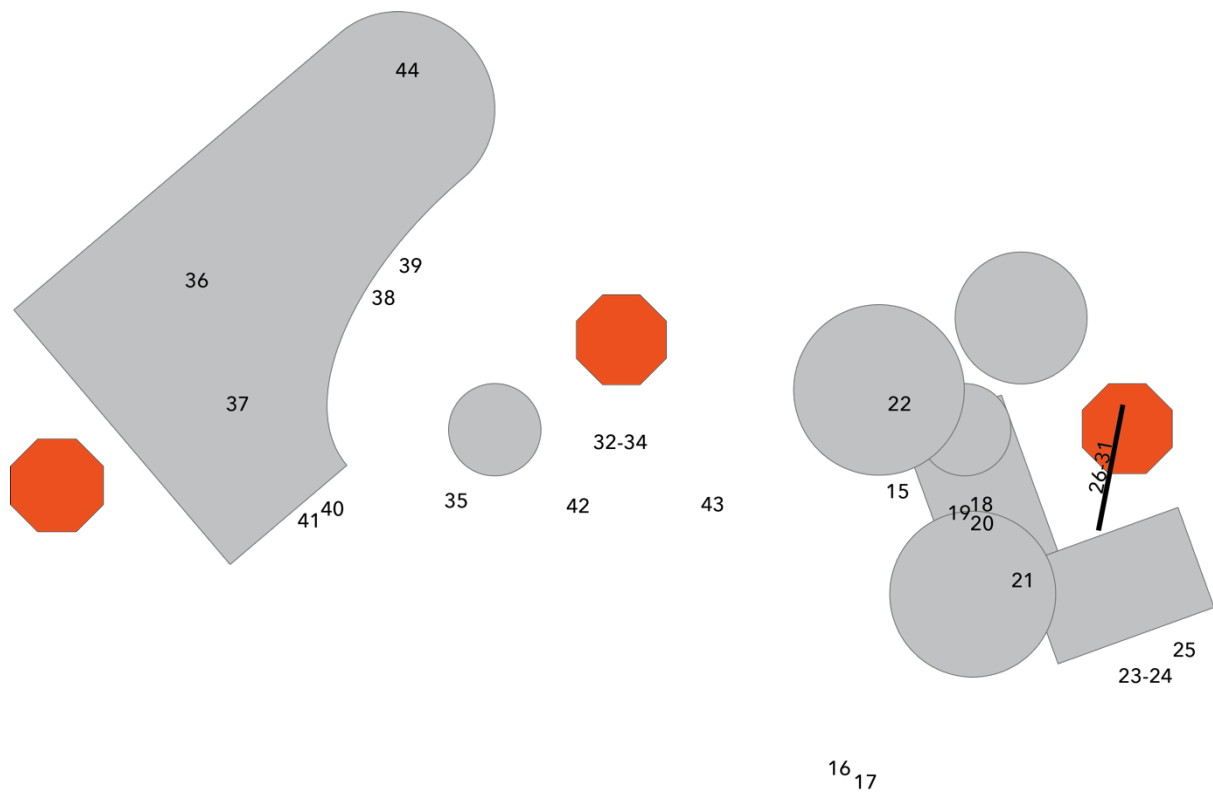


Figure 20 Approximate positions of the spot microphones with their corresponding channel numbers (own graphic)



Figure 21 DaVinci Resolve still of the recording setup and the ensemble during the recording session



## 4|3 THE RECORDING SESSION

The recording took place on December 11, 2022, between 10.30am and 4.30pm. Knowing the ensemble's preference for recording a maximum of 1-2 takes per piece, we decided to use five cameras to record the performance and to keep the same lighting setup for each piece. This provided enough video material for the edit. All of this was only possible by all participants being well-prepared, proficient, and focused.



Figure 22 View from the control booth during the recording session (own photograph)

## 4|4 MIXING APPROACH

After considering and trying various methods described in this chapter, the decision of a hybrid approach was reached, where the closeness of a spot microphone mix was to be augmented with the microphone array. While the microphone array conveyed a good spatial representation of the performance, the closeness of the spot microphones produced a far more engaging and visceral experience for the viewer/listener. This aesthetic choice was made in accordance with the ensemble and led to the decision of making a mostly static extended object bed-based mix. This meant that a 9.1.6 object bed

was used for effects while the individual instruments were summed into stereo buses panned in 3D space to facilitate stereo dynamics and spectral processing, and to minimize echoes. The 7.1.2 bed was only used for the LFE channel.

## 4 | 4 | 1 SPOT SOURCE CHOICES

The choice of microphones for the piano boiled down to the Yamahiko piezo pickup (mainly used for bass frequencies) plus the two spaced Neumann U87, as these offered the best balance between source separation and accurate sonic representation of the grand piano. The Shure SM181 above the strings proved to be too close and also posed a panning challenge. The M/S microphones, while sounding good, had too much crosstalk from the other instruments and were not suitable for the mix. The reeds were best represented with the DPA 4011 wide cardioid microphone, while it is interesting to note that during the loud passages the piano picked up most of the saxophone signal and the reeds spot microphone was only used for high frequency definition in these parts.

The drum set microphones used in the mix were the Neumann U47 FET for the kick, a Shure SM57 for the snare drum, Sennheiser MD421 for the tom, Schoeps MK22 for the glockenspiel, Neumann KM184 for the gongs, and DPA4006 and 4015 for overheads spaced equally at 30cm. Similarly to the piano, many more microphones were used to provide flexibility in post-production and to compare different microphone types. The final choice considered the sonic quality of the signals as well as the possible phase issues and crosstalk.

The modified 2L-Cube was unfortunately not suitable for this production philosophy in its completeness, but the [L], [R], [Ls] and [Rs] channels were ideal for providing spatial context in the four corners of the top layer. These signals were used in addition to the immersive reverb algorithms routed to the 9.1.6 object bed. The practice of delaying spot microphone signals to the main array was applied in order to compensate for signal delays when panned to equal distances from the listening position. The reference for this was the nearest microphone of the main array with respect to the spot microphone in question.



## 4|5 OBJECT PANNING

As we saw in the previous chapter, surround formats allow for panning around the listening position. In addition, immersive audio makes vertical panning possible, which further increases the immersive experience. The loss of precise localization by using angles above 30° for the top layer speakers (45° in Atmos) is certainly a factor worth considering. However, the effect of immersion is still strong for the same reasons as described when switching from stereo to surround: the physical reality of early reflections coming from above makes vertical panning work. The room information is now present in the playback system and is no longer projected onto the 2D plane.

The spot microphone-based instrument buses were thus also panned vertically for an immersive listening experience vastly different from a stereo or surround mix.

## 4|6 MIXING THE MODULS

Although Dolby Atmos does not offer master bus processing, there are some methods to achieve similar results. For this recording I created master faders for every object used in the session that had linked multiband compressors, equalizers and limiters. This helped to control peak levels and to achieve a homogenous mix. In addition, VCAs were used for volume automation. As mentioned before, the pieces generally did not have moving objects. There are some places where spot microphone signals move around in 3D space for a dramatic effect, but further movement proved to be distracting.

### 4|6|1 MODUL 65

Modul 65 was the first piece recorded during the session and I tried to convey the meditative nature with a homogenous and precise sound. The recurring bell and piano note are deliberately mixed in a way that melds them together and gives them equal importance. There is a considerable amount of immersive (9.1.6) reverb added to the recording to further emphasize the space created by the sparse arrangement. While the piano's bass notes, the sustained bass clarinet notes and the brushes used on the snare

drum provide closeness and warmth from the front, the rear reverb levels are increased to open up the room towards the back and increase spatiality.

## 4 | 6 | 2 MODUL 68

Modul 68 is mixed in a similar fashion as Modul 65. The dramaturgy of the piece feels less meditative and more helical in the sense that every time a musical cycle is completed the listener reaches a different level in the music's circular motion. Beginnings and endings feel ambiguous, and the closeness of the mix means to convey a guided journey through the spiral as opposed to the listener observing the piece from the outside. The shifting pulse of the kick drum accelerates and decelerates the pace of the otherwise relentless pattern.

## 4 | 6 | 3 MODUL 58

The three parts of Modul 58 are tied together by the main pattern in 7 but are all very different in nature. The opening of the piece is mixed in a way to emphasize the open crispness of the music. The prepared piano pattern drives the pulse while the drums provide sparse cymbal and kick drum sounds. The saxophone plays an open solo and ends this part by the ascending line mirrored on the piano.

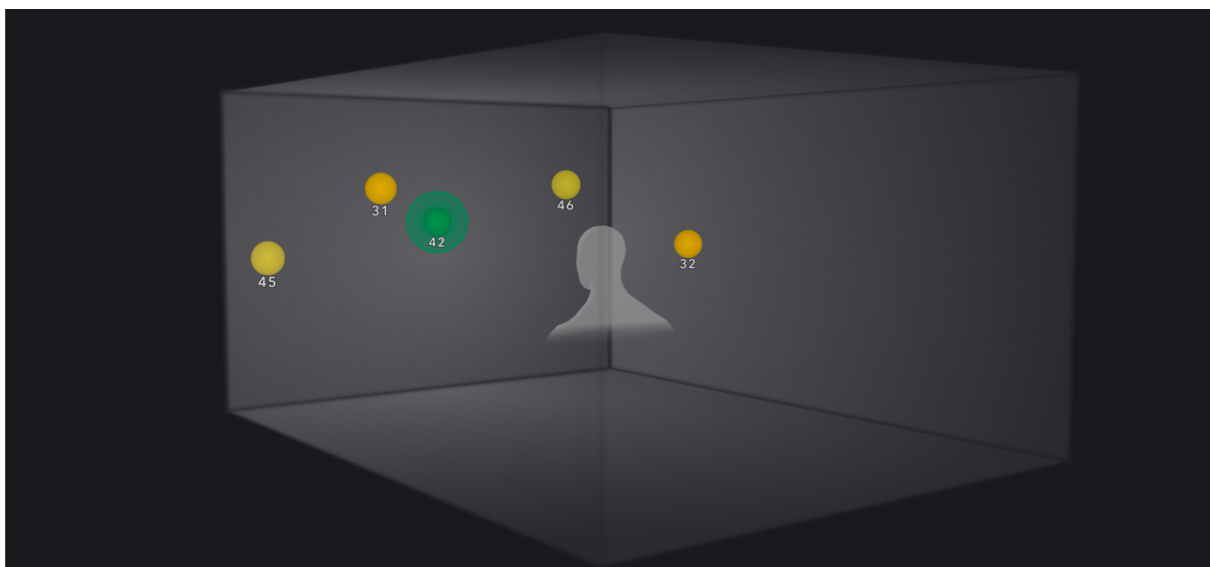
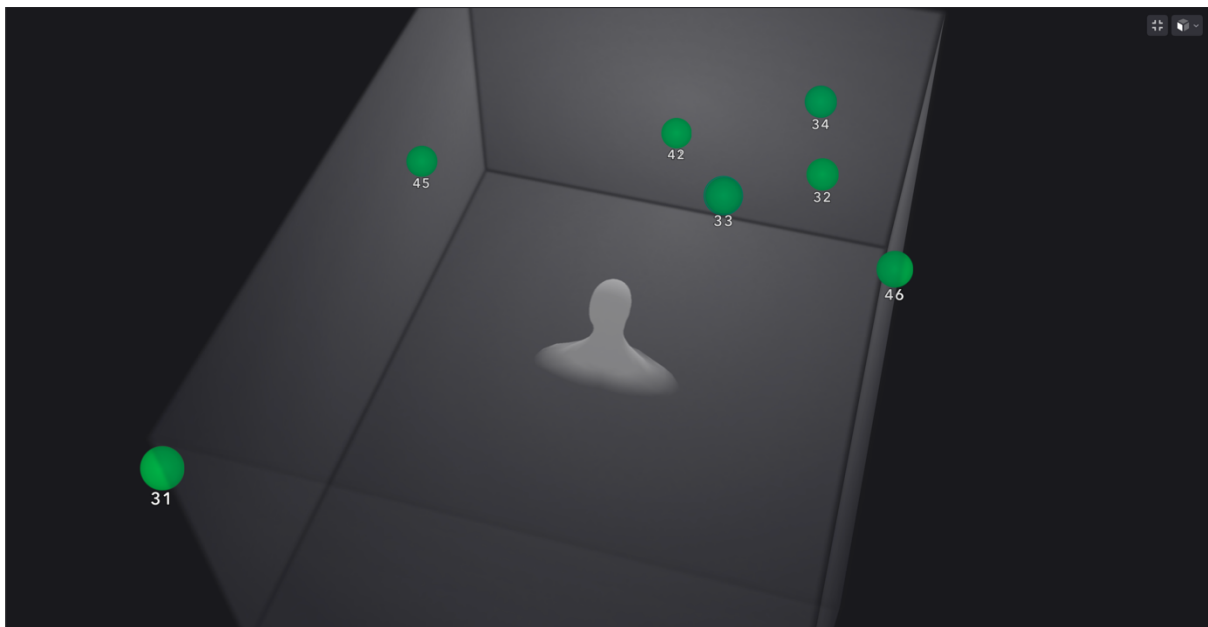


Figure 23 Screenshot of the instrument bus panning in Modul 58. 31-32 is the drum bus, 42 the bass clarinet with size 13 and 45-46 the piano bus.

The middle part with its loudness and dense complexity posed one of the main challenges while mixing due to crosstalk between the channels. Separation of the instruments was mainly done by multiband compression and equalization. The third part has significantly less reverb to draw the listener closer to the acoustic techno experience.

## 4 | 6 | 4 M O D U L 6 1

Modul 61 is the most ambient and spacious of the pieces. This mix is the most microphone array reliant (the array is about 16dB louder as in the third part of Modul 58) which adds significant height panning to the mix. The spot microphones are less prominent and can therefore be panned further apart with way less skewing of the sound field.



*Figure 24 Screenshot of the wider object panning in Modul 61. Objects 33 and 34 are the glockenspiel and gong spot microphones. As in Figure 23, the object bed is not shown for clarity.*

Modul 61 has the largest dynamic range of the recorded pieces. This was tamed by a combination of compression, limiting, and volume automation.

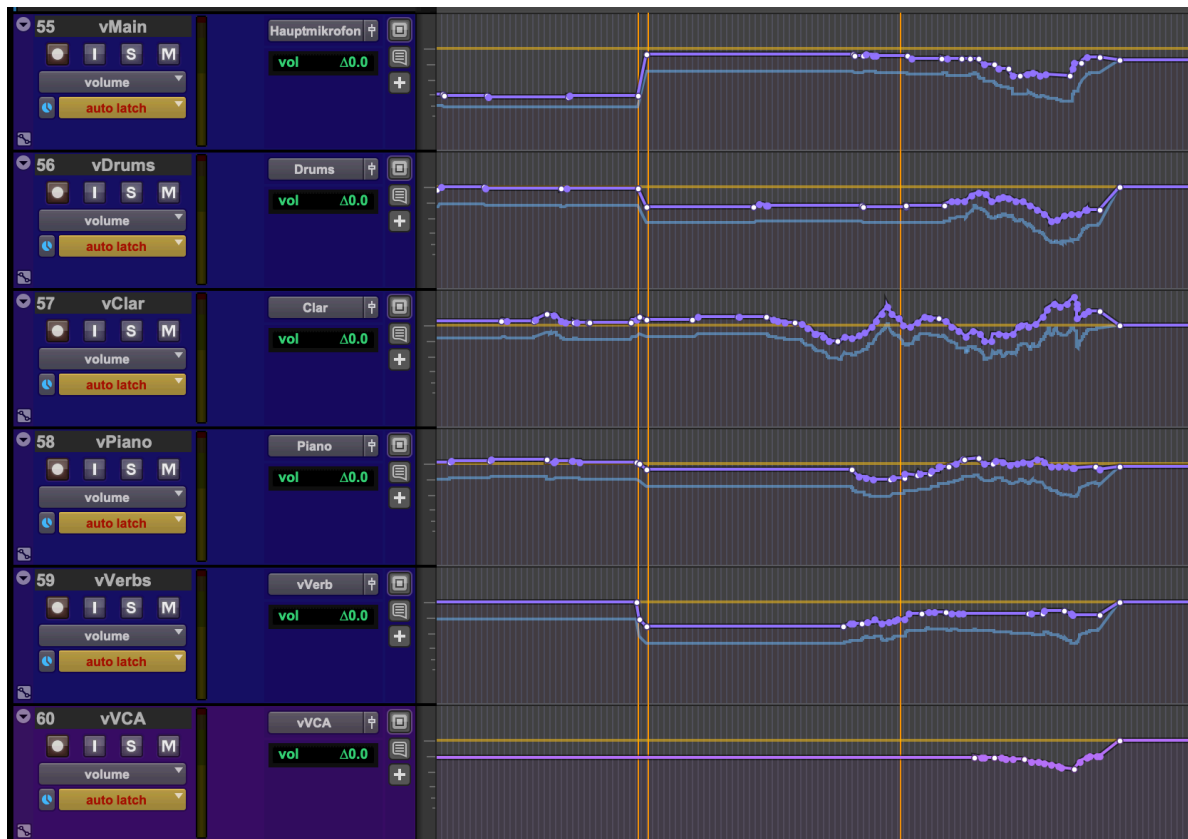


Figure 25 Pro Tools Screenshot of the VCA levels in trim mode at the end of Modul 58 and Modul 61. The level changes between pieces and in the crescendo towards the end are very apparent. The purple lines represent the individual VCA levels while the blue lines show the final VCA levels controlled by vVCA.

## 4|7 CHALLENGES

### 4|7|1 CROSSTALK

The main challenges in this production were signal cross-correlation and coherence. The setup chosen for the recording led to musical accuracy and a strong connection between the players as well as a good visual appearance for the video. This however meant that the loud passages in this dynamic music led to a lot of bleed between microphones and somewhat constrained the post-production possibilities. Firstly, equalizing the spot microphones led to equalization of more than one instrument, which made sonic compromises inevitable. Secondly, panning in 3D space was limited, as the crosstalk moved more than one instrument when moving a single instrument bus. As an example, the piano could only be moved back so far without noticeably pulling the saxophone with it. These two factors led to the mostly static object-based mix as to maintain sonic stability.

## 4 | 7 | 2 DELAY

As was shown previously, the intended playback system is an important factor to consider in Dolby Atmos. While being a scalable format in many cases, the different specifications for cinema and consumer formats can pose a significant challenge during post-production. The different loudness requirements for the end formats and the different spectral and room geometry requirements between systems means that there is no direct translation between playback systems without compromising the quality.

This became obvious when first listening to the purely object-based mix in a re-recording studio with a cinema setup. Having previously produced a satisfactory mix on a ITU setup, the differences between the systems became especially apparent when playing back transient signals. There were also resonances present that were within the more lenient Academy Curve specifications which had to do with the playback system. These were accompanied by audible echoes muddying the mix and making it unpalatable.

Spectral balancing to compensate for the different specifications of the systems is relatively easy to achieve. Solving the timing inaccuracies however is not.

While individually delaying static objects to compensate for the different times of arrival at the listening position might seem like the obvious solution, this does not translate to different venues and only works for a single listening position because of the room geometry. Dolby Atmos does take into account speaker positions of the room when utilizing objects, but they are inherently not sweet spot oriented (see 2|4|1 *Speaker Arrays*). This is of course not a problem in ITU setups, as the distances from the listener to the speakers are always equal and sufficiently small to avoid audible echoes.

Decoherent object-based signals are unproblematic in a cinematic Atmos environment, as they might arrive at slightly different times at the listener's position but usually do not significantly alter the soundscape.

Strongly coherent signals are unfortunately less suitable for object bed-based mixes in a cinematic context and require a workaround by returning to channel based mixing.

This meant that the object-based mix suitable for ITU setups was re-rendered to a 7.1.4 channel-based format. This channel count was chosen for it being the least object based mix containing top layer front to rear panning. The height channels remain static top layer objects while the horizontal plane went back to the 7.1 bed.

## 5 | CONCLUSIONS AND OUTLOOK

Dolby Atmos offers many new possibilities for creators. Being a technology rather than a philosophy it can be used in various ways and users must be aware of the benefits and drawbacks of this format.

One of the most important considerations is the target playback system of the product. There are differences between ITU-R BS.1770-1 and cinema specifications that can unfortunately become so relevant that they might require separate workflows depending on the signal content. The most obvious points being the spectral differences that make mixes sound different in these environments as well as the room geometry considerations addressed in the previous chapter.

Dolby Atmos is not directly comparable to stereo or stereo-adjacent workflows in many cases. Bus processing is very different and requires workarounds or is sometimes simply not possible. There are countless interactions between objects that cannot be directly mapped to previous formats, such as distance information as a separate attribute from volume and spectral characteristics.

It is absolutely possible to produce content in Dolby Atmos without a paradigm shift, but it is necessary to understand the differences and intricacies of the format in order to avoid production mistakes. This is especially important for loudness considerations, as different consumer formats treat loudness in very different ways that must be addressed prior to the encoding process in order to maintain quality control.

### 5|1 DECORRELATE YOUR MICROPHONE ARRAYS

Strongly cross-correlated audio content poses a special challenge to object bed-based mixes intended for cinema, as the varying room geometries and lack of time alignment can produce vastly different results in every venue. This is a physically unsurmountable factor and must be considered when choosing audio content for cinematic productions.

While less transient-heavy orchestral recordings, soundscapes and ambient field recordings might be unproblematic to use in a film mix, transient-rich cross-correlated recordings lead to inaccurate reproductions of the audio content and should be avoided or routed to beds. This principle is equally valid for musical productions as it is for sound design in film.

When recording multi-channel signals intended for cinema, decoherence is the best way to avoid timing issues during playback. As discussed in the chapter about level panning (see 3|3|2  $\Delta L$ ), real life applications require a level difference  $\Delta L > 15\text{dB}$  between two sources in order to perceive them as fully panned to one side. This means that a signal captured by one microphone in the array should ideally be more than 15dB louder than the sum of the same signal captured by the other microphones in the array.  $\Delta L = 15\text{dB}$  yields a coherence of  $\sim 0.18$ , the coherence being defined as the absolute value of the maximum of the (normalized) cross-correlation function. This amount of separation is very difficult to achieve purely with distance in real world applications, as a doubling of the distance from a source to the microphone yields  $\Delta L = 6\text{dB}$ .

Considering a distance of approximately two meters from the center channel omni microphone to the bass clarinet as in the recording done for this thesis, it would have required a distance between microphones exceeding 11m. This would have led to an array with a diameter of over 28m.

Figure 10 showed that coherence is frequency dependent. This means that signals with higher frequencies decohere faster with increasing distance than those with low frequencies. Transient signals inherently carry many frequency components and are thus more difficult to decohere properly.

Figure 11 additionally showed that microphone directivity and the angle between microphones play an important role in signal correlation. Therefore, the microphone array design for decoherent signals must take into account the nature of the sound source as well as the choice of microphone type and angles between microphones to provide the best results.

Using a main microphone array for immersive channel-based recordings remains an excellent method for capturing immersive audio, but possible complications in post-production due to signal coherence and timing differences must be taken into account. There are examples of productions using this technique that work well in a cinema environment (Morten Lindberg's recordings sound excellent) but not all content gets translated equally well between formats. The next steps would be to compare microphone arrays with different directivities to for example explore if the 2L-Cube with cardioid microphones would yield better results for a cinematic context, and to better separate spot microphones by trying to mount absorbers between them in such a closely spaced recording setup.

The mix for this thesis' presentation uses the previously described workaround of returning to bed and is a compromise. But it is the probably most realistic solution for a cinematic context. It is the ensemble's and my intention to make this recording and the video available to a wider audience in the near future, so the next steps will be to finalize the mix for consumer formats and decide on ways of distribution.

## 5|2 THE LIMITATIONS OF OBJECTS

As we saw previously, ITU playback systems are set up in a way to avoid large angles between speakers to maintain accurate phantom imaging. Surround speakers can have angles exceeding 70° and thus lose panning precision, but still create an immersive experience (see 3|5 *Postproduction Workflows in Dolby Atmos*). Speaker arrays in cinemas are designed in a way to expand the localization of the channels and thus provide a more homogenous sound reproduction, but lack the radial symmetry of ITU setups (Dolby, 2015, p. 5).

Dolby's limit of stereo-linking objects makes sense, as higher channel count based thinking is rooted in channel based workflows and leads to complications in Atmos. While it is for example possible to create phantom sources in a 7.1.4 speaker setup, even the phase interactions of coherent stereo objects can be significant. A single object can be panned in a way that its signal gets played back by 8 speakers simultaneously. This means



that even coherent stereo signals will potentially interfere on many channels and may cause undesirable comb filtering. Panning objects to discrete speaker positions eliminates electronic summing of the signals, which means that the more speakers there are in a setup, the less electronic comb filtering occurs.

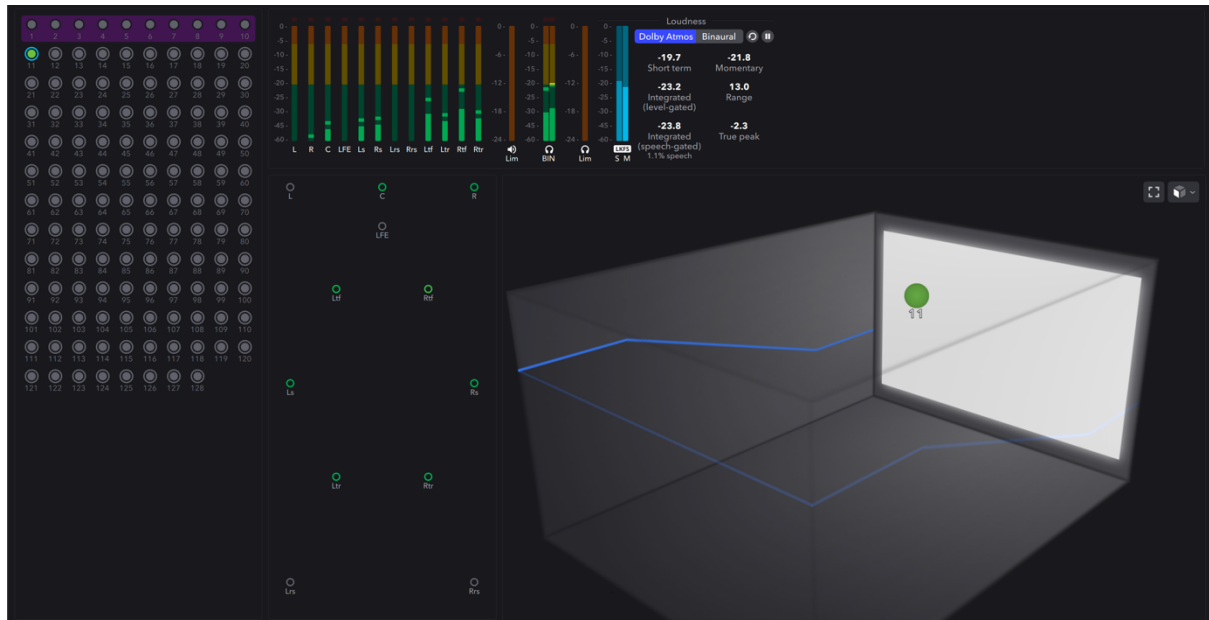


Figure 26 Screenshot of a single object panned in a way that its signal gets played back by 8 speakers in a 7.1.4 setup. The more speakers there are in a playback system, the more it can be constrained to single sound sources, lowering intrachannel coherence.

It is therefore probably the best practice to limit audio content with higher channel counts to the discrete speaker positions to minimize electronic crosstalk between playback channels.

## 5|3 OUTLOOK

At this point in time, it is unclear whether Dolby Atmos is here to stay. The endorsement by several large companies is a promising sign for this format, but it remains to be seen if the consumer market generates the demand for this technology to prevail. The possibility of proper binauralization is immensely important for its success, as the majority of content gets played back on headphones and not in calibrated studios. At the same time, label and streaming services demand mixes done in such studios, as individual binaural audio perception makes it very hard to gauge mixes properly. This requires a significant monetary investment for studio owners who want to deliver this content while there is no guarantee for Dolby Atmos to be a long-lasting format. The upcoming A-JOC (Advanced Joint Object Coding) for Dolby AC4 (Dolby, 2021, p. 11) will implement personalized HRTFs and head tracking and might prove to be the prevailing format for consumer content (Baltensperger, 2023).

One further consideration is the need for a separate stereo mix besides the Atmos mix. The downmixing algorithms in the Dolby Atmos Renderer work well but depending on the source material in the Atmos mix, signal correlations lead to a reduction of quality through the fold down process. One approach is to monitor the mix in stereo and to tweak the immersive mix until the results are adequate, but from my own experience a separate mix is oftentimes a quicker and sonically better solution.

Lastly, it is once more impressive how perception changes when multiple senses are involved. My main focus is sound design for film and this factor is an omnipresent topic when working in this field. The alignment and friction between different sensory inputs offer a powerful tool in audiovisual productions, the importance of which cannot be overstated. Our audiovisual systems work strongly on an unconscious level and can be manipulated in order to amplify or dampen effects. Just by changing the video format to a larger size, the immersive effect of this production got amplified. This led to the final choice of the aspect ratio of the recording.

## 6 | ACKNOWLEDGEMENTS

I want to thank the following people for their support during my studies and the completion of this thesis:

- Manu Gerber for his mentorship
- Nik Bärtsch, Sha and Nicolas Stocker for their friendship and their music
- Felix Scherrer and Lars Wicki for the superb edit of the video and Gaétan Nicolas for his skilled directing of photography
- Moritz Werner, Line DeKaenel, Laura Morales, Alessio Nocera, Antoinette Berta who made a fantastic recording session possible
- Roger Baltensperger for his insights into Dolby Atmos workflows
- Andreas Brüll, Andreas Werner, Daniel Hug, Olav Lervik, Tim Kleinert, Martin Scheuter and the rest of my teachers for the generous amount of new information I was so eager to soak up
- Markus Stürm and Andreas Birkle for their technical support
- Felix Baumann for helping me to safely navigate through sometimes rough seas
- My family and friends who are there to support me and my ideas. Not just in the last three years
- All cats

## 7 | BIBLIOGRAPHY

- Allen, I. (2006). The X-Curve: Its Origins and History. *SMPTE Motion Imaging Journal*, 1-24.
- Auro Technologies. (2023, July 21). Retrieved from Auro 3D: [https://www.auro-3d.com/wp-content/uploads/2022/09/Auro-3D-Home-Theater-Setup-Guidelines\\_lores.pdf](https://www.auro-3d.com/wp-content/uploads/2022/09/Auro-3D-Home-Theater-Setup-Guidelines_lores.pdf)
- Baltensperger, R. (2023, May 30). Mentorship 2023. (D. Eaton, Interviewer)
- Boenigk, O. (2010). *Einfluss der Diffusfeldkorrelation auf die räumliche Wahrnehmung bei stereofoner Wiedergabe*. Hamburg: Hochschule für Angewandte Wissenschaften Hamburg.
- Dasovich-Wilson, J., Thompson, M., & Saarikallio, S. (2022). Exploring Music Video Experiences and Their Influence on Music Perception. *Music & Science Volume 5*, 1-18.
- Di Stefano, N. (2022, August). The spatiality of sounds. From sound-source localization to musical spaces. pp. 173-185.
- Dickreiter, M., Dittel, V., Hoeg, W., & Wöhr, M. (2014). *Handbuch der Tonstudioteknik*. De Gruyter.
- Dolby. (2011). *Dolby® Surround 7.1 Technical Information for Theaters*.
- Dolby. (2015). *Dolby Atmos® Specifications*. Retrieved from Dolby: <https://professional.dolby.com/siteassets/cinema-products---documents/dolby-atmos-specifications.pdf>
- Dolby. (2021). *Dolby® AC-4: Audio delivery for next-generation entertainment services*. Dolby.
- Dolby. (2023, July 24). *What is an object?* Retrieved from Dolby Atmos Support: [https://professionalsupport.dolby.com/s/article/What-is-an-object?language=en\\_US](https://professionalsupport.dolby.com/s/article/What-is-an-object?language=en_US)
- Dolby. (2023, July 26). *What is Binaural Render Mode, and how do the settings affect my mix?* Retrieved from <https://professionalsupport.dolby.com/>: [https://professionalsupport.dolby.com/s/article/What-is-Binaural-Render-Mode-and-how-do-the-settings-affect-my-mix?language=en\\_US](https://professionalsupport.dolby.com/s/article/What-is-Binaural-Render-Mode-and-how-do-the-settings-affect-my-mix?language=en_US)

- Dolby. (2023, July 30). *Is it possible to assign different Binaural Render Mode distance models on different beds?* Retrieved from Dolby Atmos Support:  
[https://professionalsupport.dolby.com/s/article/Is-it-possible-to-assign-different-Binaural-Render-Mode-distance-models-on-different-beds?language=en\\_US](https://professionalsupport.dolby.com/s/article/Is-it-possible-to-assign-different-Binaural-Render-Mode-distance-models-on-different-beds?language=en_US)
- Dolby. (2023, July 30). *Spatial coding limitations and fine tuning*. Retrieved from  
<https://customer.dolby.com/>: <https://customer.dolby.com/content-creation-and-delivery/dolby-atmos-renderer-v500/documentation/dolby-atmos-renderer-users-guide/technology-overviews/spatial-coding/spatial-coding-limitations-and-fine-tuning>
- Dolby. (2023, July 30). *What is a bed?* Retrieved from Dolby Atmos Support:  
[https://professionalsupport.dolby.com/s/article/What-is-a-bed?language=en\\_US](https://professionalsupport.dolby.com/s/article/What-is-a-bed?language=en_US)
- Dolby. (2023, July 30). *What is spatial coding*. Retrieved from  
<https://customer.dolby.com/>: <https://customer.dolby.com/content-creation-and-delivery/dolby-atmos-renderer-v500/documentation/dolby-atmos-renderer-users-guide/technology-overviews/spatial-coding/what-is-spatial-coding>
- Dolby. (2023, July 31). *Appendix B - Room Design and System Calibration*. Retrieved from  
<https://learning.dolby.com/>:  
<https://learning.dolby.com/mod/scorm/player.php?a=144&currentorg=B0&scoid=284>
- Dolby. (2023, July 31). *Dolby Atmos for cinema playback*. Retrieved from  
<https://professional.dolby.com/>: <https://professional.dolby.com/cinema/dolby-atmos/2>
- Dolby. (2023, July 31). *Dolby Atmos Music Master Delivery Specification*. Retrieved from  
<https://professionalsupport.dolby.com/>:  
[https://dolby.my.salesforce.com/sfc/p/#700000009YuG/a/4u0000000IFJ9/IVkxI54tPvDbGymtjQ6tyljEvkaA7xPa\\_Byq6.vH\\_dA](https://dolby.my.salesforce.com/sfc/p/#700000009YuG/a/4u0000000IFJ9/IVkxI54tPvDbGymtjQ6tyljEvkaA7xPa_Byq6.vH_dA)
- Dolby. (2023, July 31). *Dolby TrueHD lossless audio*. Retrieved from  
<https://professional.dolby.com/>: <https://professional.dolby.com/tv/dolby-truehd>
- Gericke, H., & Mielke, O. (2022, March). 3D-Hauptmikrofon-Setups im Hörvergleich. *VdT Magazin*, pp. 35-39.

- Gray, J. (2023, April 11). *Mastering Dolby Atmos Music with Justin Gray - Approaches To Immersive Object-Based Mastering*. Retrieved July 2023, from Youtube:  
<https://www.youtube.com/watch?v=FcZWOCMVkWE>
- Inglis, S. (2022, July). Mixing Atmos: Morten Lindberg. *Sound on Sound Magazine*, pp. 112-117.
- ITU. (2023, July 26). *Rec. ITU-R BS.775-1*. Retrieved from ITU:  
[https://www.itu.int/dms\\_pubrec/itu-r/rec/bs/R-REC-BS.775-1-199407-S!!PDF-E.pdf](https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.775-1-199407-S!!PDF-E.pdf)
- Majidimehr, A. (2023, August 17). *Validity of X-Curve for Cinema Sound*. Retrieved from Audio Science Review:  
<https://www.audiosciencereview.com/forum/index.php?threads/validity-of-x-curve-for-cinema-sound.204/>
- Netflix. (2023, July 24). *Netflix Sound Mix Specifications & Best Practices v1.4*. Retrieved from Netflix: <https://partnerhelp.netflixstudios.com/hc/en-us/articles/360001794307-Netflix-Sound-Mix-Specifications-Best-Practices-v1-4>
- Pfanzagl-Cardone, E. (2023). *The Art and Science of 3D Audio Recording*. Springer.
- Riekehof-Böhmer, H. (2010). *Voraussage der wahrgenommenen räumlichen Breite stereofoner Mikrofonanordnungen*. VdT.
- Wikipedia. (2023, August 22). <https://en.wikipedia.org/wiki/Microphone#Cardioid>. Retrieved from wikipedia.org:  
[https://commons.wikimedia.org/wiki/File:Polar\\_pattern\\_cardioid.svg](https://commons.wikimedia.org/wiki/File:Polar_pattern_cardioid.svg)
- Wikipedia. (2023, July 30). *Spherical harmonics*. Retrieved from Wikipedia:  
[https://en.wikipedia.org/wiki/Spherical\\_harmonics](https://en.wikipedia.org/wiki/Spherical_harmonics)

## 8 | APPENDIX

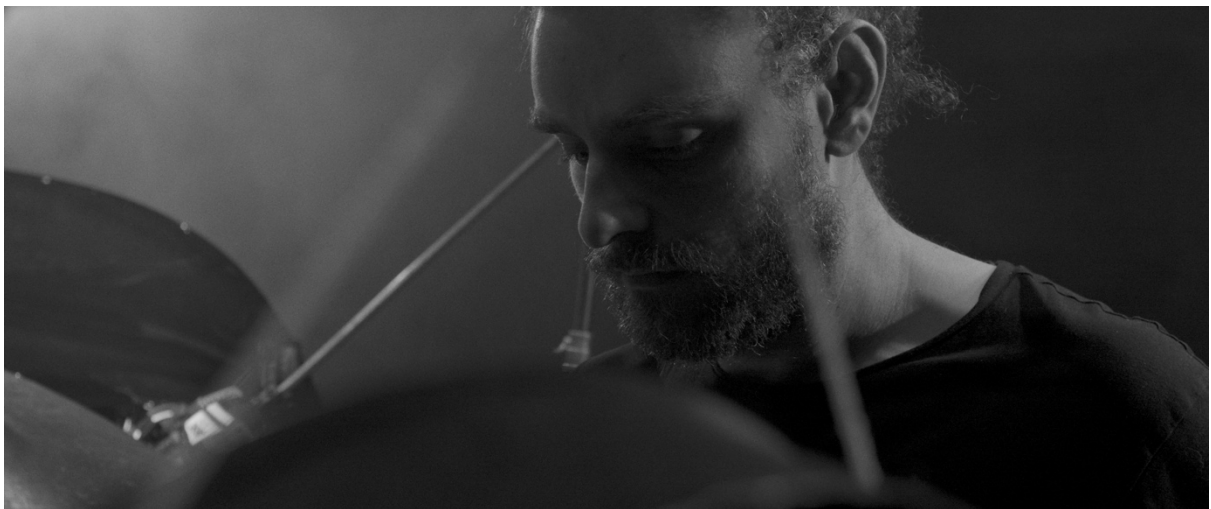
**Channel list 11.12.2022 MOBILE KS1**

Channel	Microphone	Function	Stand	Snake	Stagebox
1	DPA 4006C	L	Suspended	8.1	
2	DPA 4006C	C	Suspended	8.1	
3	DPA 4006C	R	Suspended	8.1	
4	DPA 4006C	Ls	Suspended	8.1	
5	DPA 4006C	Rs	Suspended	8.1	
6	-	-	-		
7	Neumann KM184	Top L	Suspended	2.1	
8	Neumann KM184	Top R	Suspended	2.1	
9	Neumann KM184	Top Ls	Suspended	2.2	
10	Neumann KM184	Top Rs	Suspended	2.2	
11	DPA 4006A	A	Suspended	8.1	
12	DPA 4006A	B	Suspended	8.1	
13	Neumann TLM50	Outrigger L	Suspended		
14	Neumann TLM50	Outrigger R	Suspended		
15	Neumann U47 FET	Kick Close	Short	8.2	
16	Royer R121	Kick 8	Short	8.2	
17	Schoeps MiniCMIT	Kick Shotgun	Short	8.2	
18	Shure SM57	Snare Top	Short	8.2	
19	Neumann KM84	Snare Top Fizz		8.2	
20	Shure SM57	Snare Bottom	Short	8.2	
21	Neumann KM184	Hi Hat	Tall	8.2	
22	Sennheiser MD421	Tom	Short	8.2	
23	Schoeps MK22	Glockenspiel	Tall	8.3	
24	Schoeps MK22	Glockenspiel	Tall		
25	Neumann KM184	Gongs	Tall	8.3	
26	DPA 4006A	OH L Omni	OH Boom	8.3	Sb 2 - 1
27	DPA 4006A	OH R Omni		8.3	2 - 2
28	DPA 4011A	OH L Cardioid		8.3	2 - 3
29	DPA 4011A	OH R Cardioid		8.3	2 - 4
30	Neumann KM184	OH X		8.3	2 - 5
31	Neumann KM184	OH Y		8.3	2 - 6
32	DPA 4015C	Clar High	Short		2 - 7
33	Neumann U67	Clar Low	Short		2 - 8
34	DPA 4006C	Clar Omni	Short		2 - 9
35	Neumann KM184	Tom	Short		2 - 10
36	Shure SM181	Piano in Low	-	8.4	2 - 11
37	Shure SM181	Piano in High	-	8.4	2 - 12
38	Neumann U87	Piano close L	Tall	8.4	2 - 13
39	Neumann U87	Piano close R		8.4	2 - 14
40	DPA 4006C	Piano M	Tall	8.4	2 - 17
41	Schoeps CCM8	Piano S		8.4	2 - 18
42	Schoeps BLM 03C	Floor Sha	-		2 - 21
43	Schoeps BLM 03C	Floor Nicolas	-		2 - 22

*Table 1 Complete channel list of the recording session*



*Figure 27 DaVinci Resolve still of Nik Bärtsch*



*Figure 28 DaVinci Resolve still of Nicolas Stocker*



*Figure 29 DaVinci Resolve still of Sha*



## 9 | STATEMENT OF ACADEMIC HONESTY

*I hereby state that I authored this thesis and produced the media mentioned without external assistance. All passages that contain quotes or content from external sources are declared and referenced.*

A handwritten signature in black ink, appearing to read 'd. Eaton', with a stylized horizontal line extending to the right.

*Daniel Eaton, Zürich, August 2023*