

# ICSA 2019

## Proceedings of the

5th International Conference on Spatial Audio

September 26th to 28th, 2019

Ilmenau, Germany

Edited by

Stephan Werner and Steve Göring





# Audio for Virtual, Augmented and Mixed Realities

Proceedings of ICOSA 2019

5th International Conference on Spatial Audio

September 26th to 28th, 2019

Ilmenau, Germany

**Edited by**

Stephan Werner and Steve Göring

**ilmedia**

2019

# Imprint

Audio for Virtual, Augmented and Mixed Realities  
Proceedings of ICSA 2019

## Edited by:

Stephan Werner  
Technische Universität Ilmenau  
Department of Electrical Engineering and  
Information Technology  
Electronic Media Technology Group  
P.O. Box 10 05 65  
98684 Ilmenau  
Germany

Steve Göring  
Technische Universität Ilmenau  
Department of Electrical Engineering and  
Information Technology  
Audio-Visual Technology Group  
P.O. Box 10 05 65  
98684 Ilmenau  
Germany

## Organized by:

Verband Deutscher Tonmeister e.V.  
Am Zaarshäuschen 9  
51427 Bergisch Gladbach  
Germany  
<https://tonmeister.org/>

Technische Universität Ilmenau  
Department of Electrical Engineering and  
Information Technology  
Electronic Media Technology Group  
P.O. Box 10 05 65  
98684 Ilmenau  
Germany  
<https://www.tu-ilmenau.de/en/institute-for-media-technology/>

## Publisher:

Universitätsbibliothek Ilmenau  
[ilmedia](#)  
P.O. Box 10 05 65  
98684 Ilmenau  
Germany

**DOI:** 10.22032/dbt.39936  
**URN:** urn:nbn:de:gbv:ilm1-2019200492

© Verband Deutscher Tonmeister e.V., 2019

## Editorial

The ICSA 2019 focuses on a multidisciplinary bringing together of developers, scientists, users, and content creators of and for spatial audio systems and services. A special focus is on audio for so-called virtual, augmented, and mixed realities.

The fields of ICSA 2019 are:

- Development and scientific investigation of technical systems and services for spatial audio recording, processing and reproduction
- Creation of content for reproduction via spatial audio systems and services
- Use and application of spatial audio systems and content presentation services
- Media impact of content and spatial audio systems and services from the point of view of media science.

The ICSA 2019 is organized by VDT and TU Ilmenau with support of Fraunhofer Institute for Digital Media Technology IDMT.

## Hosts



## Support Partners





## Index

<b>Editorial</b> .....	iii
Mäsel, Jakob; Simon, Christian; Choi, Jin; Schulz, Andreas; Silzle, Andreas .....	1
Field Test for Immersive and Interactive Audio Production with the Gewandhausorchester Leipzig using MPEG-H	
DOI: 10.22032/dbt.39945	
Grab, Janet .....	9
Capturing 3D Audio: A pilot study on the spatial and timbral auditory perception of 3D recordings using main-array and front-rear separation in diffuse field conditions	
DOI: 10.22032/dbt.39946	
Arlauskas, Gabriel; Ohland, Jonas; Schaar, Henning.....	17
Ambisonics Decoder Description (.ADD) : developing a new file format for Ambisonics decoding matrices	
DOI: 10.22032/dbt.39947	
Heidtmann, Stefan; Fohl, Wolfgang.....	21
A Systematic Performance Investigation of Convolution Algorithms for Synthetic Room Reverberation	
DOI: 10.22032/dbt.39948	
Dodds, Peter; Amengual Garí, Sebastià V.; Brimijoin, W. Owen; Robinson, Philip W.....	29
Auralization systems for simulation of augmented reality experiences in virtual environments	
DOI: 10.22032/dbt.39949	
Fehling, Matthias; Nogalski, Malte; Wilk, Eva.....	35
Implementation of Ambisonics Recordings in a Wave Field Synthesis System	
DOI: 10.22032/dbt.39950	
Devonport, Sean; Foss, Richard .....	39
The Distribution of Ambisonic and Point Source Rendering to Ethernet AVB Speakers	
DOI: 10.22032/dbt.39951	

Pörschmann, Christoph; Arend, Johannes Mathias .....	47
How positioning inaccuracies influence the spatial upsampling of sparse head-related transfer function sets	
DOI: 10.22032/dbt.39952	
Rummukainen, Olli S.; Robotham, Thomas; Plinge, Axel; Wefers, Frank; Herre, Juergen; Habets, Emanuël A. P. ....	55
Listening Tests with Individual versus Generic Head-Related Transfer Functions in Six-Degrees-of-Freedom Virtual Reality	
DOI: 10.22032/dbt.39954	
Plinge, Axel; Schlecht, Sebastian; Rummukainen, Olli; Habets, Emanuël A. P. ....	63
Enhanced Immersion for Binaural Audio Reproduction of Ambisonics in Six-Degrees-of-Freedom: The Effect of Added Distance Information	
DOI: 10.22032/dbt.39955	
Robotham, Thomas; Rummukainen, Olli S.; Habets, Emanuël A. P. ....	71
Towards the Perception of Sound Source Directivity Inside Six-Degrees-of-Freedom Virtual Reality	
DOI: 10.22032/dbt.39956	
Hestermann, Simon; Lukashevich, Hanna; Sladeczek, Christoph .....	79
Deep Neural Network Approaches for Selective Hearing based on Spatial Data Simulation	
DOI: 10.22032/dbt.39957	
Weidner, Florian; Fiedler, Bernhard; Redlich, Johannes; Broll, Wolfgang .....	85
Exploring Audiovisual Support Systems for In-Car Multiparty Conferencing	
DOI: 10.22032/dbt.39958	
Beer, Daniel; Brocks, Tobias; Küller, Jan; Strehle, Steffen; Koch, Tilman.....	93
New potentials for portable spatial audio with MEMS based speakers	
DOI: 10.22032/dbt.39959	
Cairns, Patrick; Moore, David.....	99
Switched Spatial Impulse Response Convolution as an Ambisonic Distance-Panning Function	
DOI: 10.22032/dbt.39961	

Resch, Thomas; Hädrich, Markus .....	107
The Virtual Acoustic Spaces Unity Spatializer with custom head tracker	
DOI: 10.22032/dbt.39962	
El Baba, Youssef; Walther, Andreas; Habets, Emanuël.....	115
Room geometry inference using sources and receivers on a uniform linear array	
DOI: 10.22032/dbt.39963	
Schultz, Frank; Hahn, Nara; Spors, Sascha .....	123
Detection of Constant Phase Distortions in Filters for Sound Field Synthesis	
DOI: 10.22032/dbt.39964	
Frank, Matthias; Brandner, Manuel.....	131
Perceptual Evaluation of Spatial Resolution in Directivity Patterns 2: coincident source/listener positions	
DOI: 10.22032/dbt.39966	
Werner, Stephan; Klein, Florian; Götz, Georg.....	137
Investigation on spatial auditory perception using non-uniform spatial distribution of binaural room impulse responses	
DOI: 10.22032/dbt.39967	
Klein, Florian; Neidhardt, Annika; Seipel, Marius .....	145
Real-time Estimation of Reverberation Time for Selection of suitable binaural room impulse responses	
DOI: 10.22032/dbt.39968	
Blochberger, Matthias; Zotter, Franz; Frank, Matthias .....	151
Sweet area size for the envelopment of a recursive and a non-recursive diffuseness rendering approach	
DOI: 10.22032/dbt.39969	
Genovese, Andrea; Gospodarek, Marta; Roginska, Agnieszka.....	159
Mixed Realities: a live collaborative musical performance	
DOI: 10.22032/dbt.39971	

Neidhardt, Annika .....	165
Data set: BRIRs for position-dynamic binaural synthesis measured in two rooms	
DOI: 10.22032/dbt.39972	
Nipkow, Lasse .....	171
3D audio for live events	
DOI: 10.22032/dbt.39973	
Ott, Johannes; Wienböcker, Niklas; Tutescu, Anca-Stefania; Rosenbauer, Jan; Görne, Thomas .....	179
Spatial audio production for immersive fulldome projections	
DOI: 10.22032/dbt.39974	





## Abstract Reviewed Paper at ICSA 2019

Presented by VDT.

### Field Test for Immersive and Interactive Audio Production with the Gewandhausorchester Leipzig using MPEG-H

Jakob Mäsel<sup>1</sup>, Christian Simon<sup>1</sup>, Jin Choi<sup>2</sup>, Andreas Schulz<sup>3</sup>, Andreas Silzle<sup>1</sup>

<sup>1</sup> *Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany, Email: jakob.maesel @iis.fraunhofer.de*

<sup>2</sup> *Sempre La Musica Korea, Seoul, Republic of Korea*

<sup>3</sup> *Gewandhaus zu Leipzig, Leipzig, Germany*

#### Abstract

Next Generation Audio (NGA) codecs offer new features for consumers such as advanced user interactivity, immersive sound and optimized reproduction across different classes of playback devices. Yet, 'real world' experience in the application of NGA workflows in classical music production is limited: The realistic depiction of the original concert hall sound and the treatment of elevated sound sources impose challenges when recording and mixing. Besides producing for immersive reproduction systems, sound engineers face new tasks like the integration of user interactivity, ensuring downmix compatibility and loudness consistency.

This paper expounds our field test for an NGA production workflow found suitable for complex orchestral music. Using MPEG-H 3D Audio as an example, we recorded three full-length concerts with Gewandhausorchester Leipzig. Among them is the prestigious new year's concert with Beethoven's 9<sup>th</sup> symphony conducted by Gewandhauskapellmeister Andris Nelsons. We developed a microphone setup for the immersive recording of classical music in Gewandhaus and achieved convincing results with the recording and mixing strategies presented in this paper. The perceived reverberance in the mixes was found to be realistically relatable to the excellent room sound in the original concert hall. Elevated sources, like the choir and soloists, were clearly depicted as such. For metadata authoring and mastering, MPEG-H production tools were used to regulate loudness, dynamic range and downmix compatibility, targeting playback over loudspeakers and 3D audio soundbars. Additionally, we tested two different approaches for interactivity that are suitable for classical music listeners.

Key words: Next Generation Audio, MPEG-H Audio, immersive music production, user interactivity, classical music.

## 1 Introduction

Next Generation Audio (NGA) codecs provide the foundation for producing, distributing and playing back immersive and user-interactive content, whether it be channel- or object-based audio. The ISO standard MPEG-H 3D Audio<sup>1</sup> [1] allows for transmitting and reproducing immersive mixes, user interactivity or accessibility features. It is further possible to render content to different reproduction systems with varying channel configuration. This makes it easy to serve different playback scenarios with only one dedicated mix: Whether it be stereo, surround, immersive formats with elevated speakers or binaural reproduction [2].

Although all the necessary tools exist, there are only few immersive music recordings fully exploiting the new features of NGA. There are even less publications on real world applications, especially in classical music. Yet the upcoming of 3D soundbars opens up the mass market towards the reproduction of NGA content.

Recording for immersive playback scenarios with elevated sound sources raises new questions and there is hardly any publication on end to end chains for immersive and user-interactive music productions. Thus, with our field test, we aim to find an NGA production workflow that is suitable for complex orchestral music. This paper documents our immersive recording of the Gewandhausorchester's New Year's concert 2018/2019 and the according preliminary recording sessions. Our main target was the realistic depiction of the listener's perception of the concert hall, maintaining the high standards of today's classical music production. Besides room accuracy, we focused on the reproduction of elevated sources, such as choir and soloists, as they were placed on a tribune behind the orchestra.

Using the example of our immersive recordings in Gewandhaus (see Fig. 1) and the according mixing process, this paper will further delineate the opportunities of mastering object based audio (OBA) during the authoring process. This includes handling loudness and downmixing in the course of universal delivery. Subsequently, two possible scenarios for user interactivity are exemplified in section 4.3.

## 2 Discussion of the Chosen Production Setup

Understanding the principles behind multichannel recording of classical music for surround reproduction setups is the basis for recording in 3D. Theile [3] presented different approaches, substantiated by ORF (Austrian public service broadcasting) listening tests conducted by Camerer [4]. They compared many different surround microphone arrays and reported outstanding image stability and good evaluation of timbre for at least the Optimized Cardioid Triangle (OCT) and the Decca Tree respectively. Thus, in Gewandhaus we compared OCT 3D and Decca Tree plus wide spaced microphones and the Hamasaki square for surround and height layer according to the findings of Hamasaki et al. [5]



**Fig. 1:** 3D audio recordings were performed in Gewandhaus zu Leipzig, Germany.

[6]. To obtain flexibility for adding more diffuse sound compared to the main microphone system, we set up a more distant wide spaced AB microphone setup. The combination of Decca and AB is well known and has a long history in famous recording studios [7]. Its widely observed preference over the spatially more accurate OCT system for symphonic orchestra content could be explained with the use of omnidirectional microphones in AB/Decca arrays instead of cardioid or super-cardioid microphones in OCT. This provides a more linear bass depiction and is thus said to draw a more realistic picture in terms of timbre [8]. This coincides with the studies of Rumsey et al. [9] on weighted preference: They found that timbral fidelity is a dominant influence on overall sound quality ratings with an impact of 67% for surround sound reproduction. It is followed by front localization with 25% and surround localization (8%). This suggests that localization accuracy might not be the top priority for main microphone array design although it still plays a considerable role. It is surely interesting how the weighted preferences change when introducing envelopment and engulfment as new independent parameters, strongly depending on the use of horizontal and vertical loudspeakers respectively [10].

The principles of spatial and directional depiction in surround recording also apply to 3D production techniques. They are well known but require careful setup adjustment and experimentation [11]. Yet, choosing microphone setups for 3D recording imposes new challenges due to the increasing amount of channels and their according relationships: With more channels, interchannel crosstalk is more likely to happen and more difficult to avoid [12]. This coincides with studies by Scuda [13] and Grewe [14]: For the evaluation of microphone setups, they reported strong dependencies on musical content and recording location. Therefore we used and combined different main microphone setups in order to stay flexible regarding different musical material and instrumentation.

Field tests for recording and transmitting 3D audio content have been publicized rarely but on a regular basis: Nishiguchi et al. [15] from Japanese broadcaster NHK were one of the first ones reporting about it for their 9+10+3<sup>2</sup> (22.2) speaker

<sup>1</sup> Further referred to as MPEG-H

<sup>2</sup> ITU-R BS 2051 notation (Height+Mid+Lower)

configuration. Field tests in sports broadcasting using MPEG-H were conducted by Stenzel et al. [16]. The EU research project ORPHEUS evaluated in detail all aspects of a complete end-to-end OBA chain for audio-only content [17]. In corporation with the European Broadcast Union (EBU) and major European broadcasting stations, they successfully showed that the production and transmission of object-based and user-interactive 3D audio content is possible with the NGA codec MPEG-H, explaining the full chain and workflow [18] including extensive user experience tests [19]. Many ideas for content were created and tested during their experiments. Simon et al. conducted field tests for the immersive and user-interactive live production and transmission of the Eurovision Song Contest 2018 and French Tennis Open 2018 [20]. The field tests outlined in this paper extend the previous described activities to the field of classical music.

### 3 Production and Distribution of Classical Music with MPEG-H

NGA enables a whole new way to produce and consume audio. Legacy codecs are only able to deliver unchangeable mixes. With OBA, mixes can adapt to the capabilities of their respective playback device and be personalized by consumers. This is realized by the production and delivery of discrete audio components, which are then flexibly rendered in the playback device with the help of affixed metadata. The new process of the production workflow is called authoring, creating a so-called audio scene, including all audio and metadata.

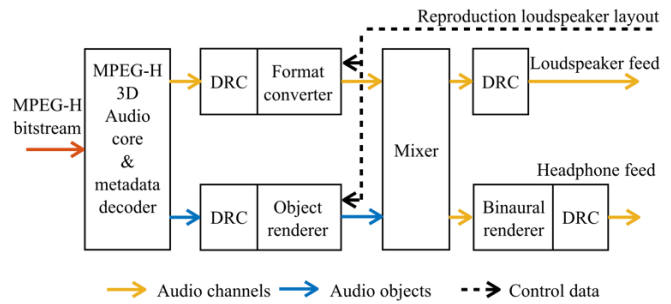
MPEG-H comprises three NGA features, which are all relevant for classical music production:

- Immersive Audio
- Interactivity
- Universal Delivery

Immersive audio refers to three-dimensional sound and enhanced envelopment compared to speaker configurations that are not using the height layer. It is possible to produce for certain 3D audio channel layouts defined in [1] as well as to create layout-independent audio objects defined by their relative position to the listener.

Interactivity allows for changing the mix in the playback device. Producers decide which parameters the listeners can adjust and to what extent. This can be selecting presets with different presentations of the delivered audio components. Even changes in the position of objects or their relative volume levels are conceivably possible.

Universal delivery enables for playback of the same audio content in a wide range of different reproduction scenarios. It comprises a format converter and an elaborate dynamic range control (DRC). The format converter changes the raw multi-channel audio stream to the selected playback format, as can be seen in Fig. 2. This concept allows for producing only in the largest reproduction format, while the adaptation for smaller formats happens in the renderer of the playback device. The converter uses an active downmix algorithm to avoid downmix artefacts and ensure artistic integrity. The downmix



**Fig. 2:** Top-level block diagram of MPEG-H 3D Audio decoder with dynamic range control (DRC). All depicted components are controlled by metadata.

coefficients in MPEG-H are highly flexible and can be defined by the producer.

The DRC adapts the overall level to a device-dependent target loudness and reduces the dynamic range accordingly. It is based on the integrated loudness measurement of each component contained in the audio scene. In [1], three target loudness profiles are defined:

- Audio Video Receiver (AVR): -31 LUFS
- TV: -24 LUFS
- Mobile devices: -16 LUFS

Furthermore, the MPEG-H system automatically normalizes program loudness according to common standards (e.g. EBU R-128 [21] [22], ITU-R BS.1770.4 [23], ATSC A/85 [24], etc.) providing loudness consistency between different content. Also, the loudness between different presets is adapted, avoiding leaps in level during user interaction. In certain cases, especially in OBA mastering for music, it can be necessary to replace the automatic loudness adaption with manual loudness settings, for example to preserve the natural or aesthetically motivated loudness relations between single movements. All mentioned features are under full control of producer or broadcaster.

Dedicated production tools enable for authoring MPEG-H content: They provide the mentioned features, allow for monitoring of interactivity and universal delivery and, eventually, export audio and affixed metadata. The export format can be BWF-ADM [25] or the MPEG-H Production Format (MPF), consisting of the audio data and metadata, the latter modulated to an audio file or stream in the so called Control Track [26].

## 4 Field Test

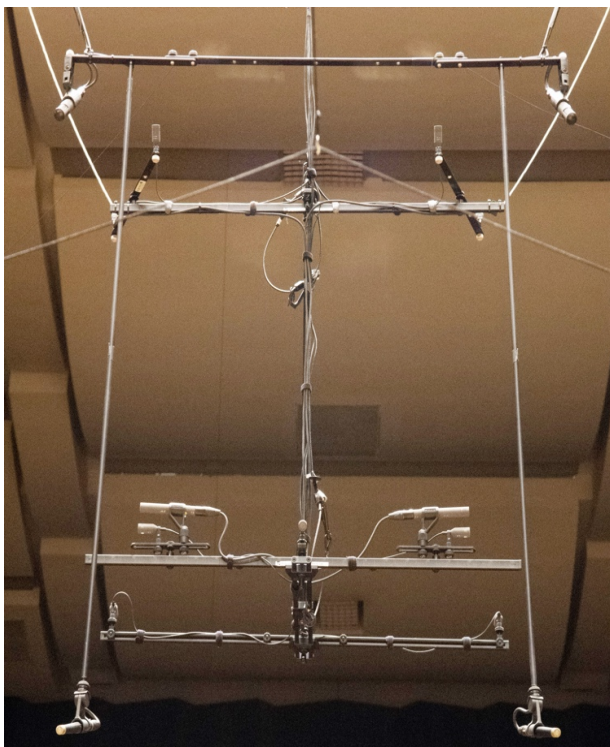
### 4.1. Recording and Mixing

Over the course of three different concerts, we experimented with different microphone setups. We focused on symphonic material with elevated sound sources, like choir and vocal soloists, as the third and main recording was supposed to be Beethoven's 9<sup>th</sup> Symphony played by the Gewandhausorchester and conducted by Gewandhauskapellmeister Andris Nelsons.

As main microphone we compared three different setups (see Fig. 3):

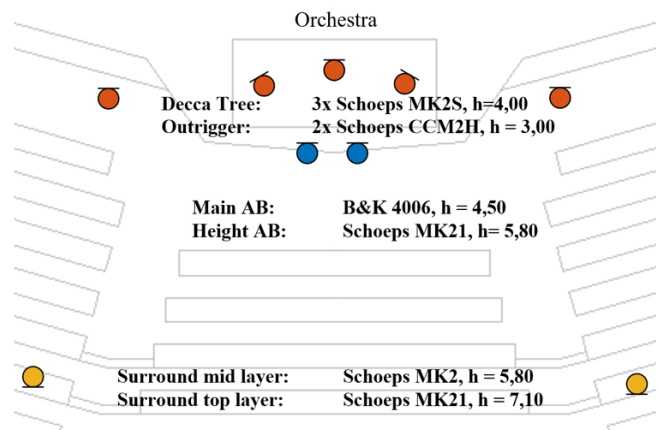
- (1) OCT 3D with additional omnidirectional microphones on front left and right to compensate for the restrained bass response of the hyper-cardioid microphones
- (2) AB with omnidirectional microphones, spaced 1 meter
- (3) Decca Tree with omnidirectional microphones, spaced 2.20 meters, with retracted center (30 cm ahead)<sup>3</sup> and outriggers

The mixing was done on a 4+5+0 loudspeaker setup in the so called “Mozart” studio at Fraunhofer IIS audio laboratories, [27]. During the mixing sessions, we subjectively evaluated the setups and found the OCT to have the most accurate localization and the benefit of very controllable elevation in the 3D mix due to good signal separation. On the other hand, poor low frequency fidelity was perceived due to the use of super cardioids. Adding the signal of the additional omnidirectional microphones helped soften this impression. Still, the OCT’s stereo image was perceived more narrow and center-focused compared to all other setups. This was becoming more obvious during playback on a 3D soundbar: the OCT direct signal was found to stay mostly within the soundbar’s geometrical limits, while spaced omnidirectional setups appeared broader than the soundbar itself. The spaced omnidirectional setup enhanced the feeling of envelopment.



**Fig. 3:** Experimental main microphone setup in Gewandhaus zu Leipzig with: spaced height AB (Sennheiser MKH 800 Twin); OCT 3D height layer, OCT front with additional omnidirectional microphones on front left / front right, surround; and spaced AB (top to bottom).

<sup>3</sup> This approach is similar to previous surround recordings of Polyhymnia International.



**Fig. 4:** Final main microphone setup for the 3D recording of Beethoven’s 9<sup>th</sup> Symphony in Gewandhaus zu Leipzig: Decca Tree with Outriggers; spaced AB and height AB; mid and top layer surround microphones (from top to bottom). All designations in meter.

Furthermore, it produced a distinct string sound that could neither be achieved with the OCT nor its combination with omnidirectional microphones. As we were recording string-focused symphonic content, we eventually decided on a combination of Decca Tree and a more diffuse AB system to gain flexibility in the mix (see Fig. 4). It is important to point out that this is not a general recommendation, as the choice of main microphone setups remains a rather subjective decision depending on the recording location, material and artistic taste. This being said, in the very same concert hall we would have opted for the OCT setup for music with many individually localizable elements. Our decision was based on the musical content not least, suggesting the blending of the orchestral instruments and a very broad envelopment to support the composer’s intention: the fusion “of the worldly and the sacred” [28].

Using omnidirectional microphones for the height layer bears the risk of overly prominent direct sound in the height channels [29], which could lead to vertical blur in localization and tone coloration after downmixing. Thus, in vertically spaced arrays Lee et al. [30] suggest directional polar patterns, angled outwards to minimize interchannel crosstalk. For experimentation purposes we therefore used MKH 800 twin microphones to stay flexible regarding the degree of directivity as well as a combination of omnidirectional and cardioid microphones. In the final setup we used two Schoeps MK21, placed 1,30 meters above the main AB with same base width. They were angled up- and outwards to provide sufficient separation from the orchestra’s direct sound and still catch enough direct sound from the elevated sources to situate them in the height layer. We used the Hamasaki Square as additional room system feeding the height channels.

In the surround microphones we were facing disturbingly hard reflections from the concrete balustrade. The choice of microphone type, position and angle overcomes this



challenge. The surround's height layer is designed in the exact same way as the front height: 1,30 meters above the according mid layer microphone, angled up- and outwards. They are widely spaced (distance L-R: 12 meters) in order to increase interaural fluctuations in the low frequencies and thus envelopment [11].

The concerts were monitored and recorded in a special 3D audio OB Van from B&R Media, allowing loudspeaker playback of immersive content. It was also possible to listen to the active MPEG-H downmixes and a binauralized version of the recording in order to evaluate downmix compatibility in an early stage.

In our mix, the main height system played a distinct role, not only increasing vertical spread and the perceived size of room and orchestra but also providing subjectively satisfying musical balance and fusing of voices in the choir. Yet, regarding the apparent elevation of sources, their spot microphones were the decisive contributors. The very narrow polar pattern of Microtech Gefell's KEM microphones, used as spots for the elevated sources, was useful to isolate the choir from the instruments and make sure that the orchestra sound does not bleed into the height layer when enhancing the choir's direct sound. We delayed these spots in accordance to their measured distance from the main system to keep its sonic blending and spectral balance and filtered them to reduce the apparent proximity. Spots used to elevate sources were not panned to intermediate positions between loudspeakers. So they were panned directly to the height L and R speaker position. In order to ensure downmix compatibility, they were distributed to one respective channel. Further, we used two surround reverberators; one for the height and one for the mid layer to create reflection patterns for the spot microphones used.

Subtle denoising was applied in the pre-mastering process, yet we did not remove coughs and audience noise because, especially during silent parts, they give a good impression about both, the size and quality of the concert hall and the playback system.

## 4.2. Mastering and Authoring in MPEG-H

The necessary steps in the MPEG-H mastering and authoring process for the production described in this paper were:

- (1) Configuration of channel layout
- (2) Configuration of contained objects
- (3) Configuration of presets
- (4) Configuration of user-interactivity
- (5) Configuration of loudness values
- (6) Modification of downmix coefficients
- (7) Labeling in different languages
- (8) Audio and metadata export

We used the MPEG-H Authoring Plugin (MHAPI) [31] for authoring the final mix in the DAW mastering session. In MHAPI, audio is organized in two layers: Components and presets. A component represents an incoming mono or multichannel audio track which is affixed with metadata like labels and interactivity ranges. Presets contain one or more

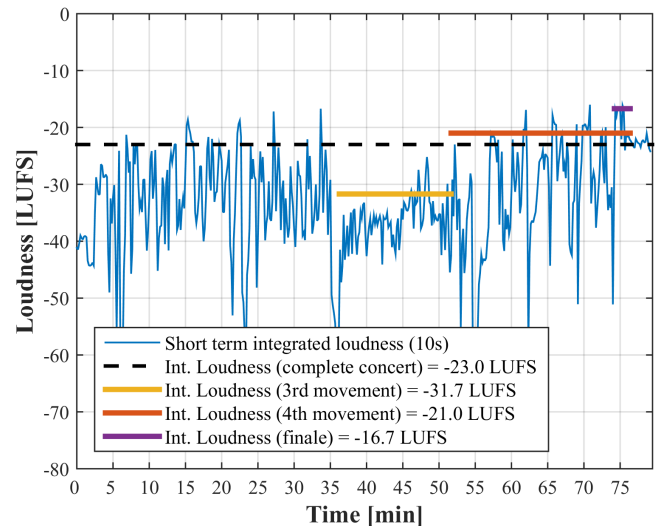


Fig. 5: Short term integrated loudness over the whole concert recorded. The overall program loudness was adjusted to -23.0 LUFS (dotted line), the loudness values of further parts of the symphony are shown.

components in different compilations and/or with different possibilities of interactivity. They allow the user to easily switch between different versions of the audio scene, but it is also possible to only author one preset.

For simplicity's sake, the following delineates the procedure for authoring audio content with only one preset without interactivity: First, the components have to be created. We chose a 4+5+0 channel format as target layout and assigned the respective input channels in the right order. No further components needed to be defined. Next, we created a preset and included the 4+5+0 audio component. As this is the only preset, it acts as default preset, active whenever the item is played back.

As explained in chapter 3, MPEG-H comprises a program loudness measurement and a loudness normalization based on it. Different movements or even parts of movements, like the finale marked purple in Fig. 5, have different loudness values. In order to preserve the artistic intention, these relative loudness differences must be maintained when exporting singular items from one cohesive program. After exporting e.g. the quiet third movement (yellow) solely, decoders would raise the item's level to match their respective target level, e.g. -23 LUFS for TV reproduction. The much louder fourth movement, depicted in red, would be attenuated to the same target level. Thus the relative loudness difference between the movements would be lost, resulting in a flattening of the symphony's dynamic that diametrically opposes the musical intent. To overcome this challenge, *album loudness* values can be set [32]. In this example, for each item being exported, its measured loudness was replaced with the measured overall program loudness, so decoder level adaptations do not influence the relative loudness differences within a compilation of individual items.

To ensure compatibility to reproduction systems with fewer channels, downmix rules can be defined by producers. In the case of the production described here, it was important to

preserve the direct sound of the elevated sources, choir and soloists, which were exclusively panned to the top front speakers in the 4+5+0 mix. Thus, those channels were only slightly attenuated for the 0+5+0 and 0+2+0 downmix configurations. On the other hand, surround and top surround objects were attenuated more, in order to avoid overly diffuse downmixes. The downmixes can be monitored in MHAPI, allowing control over all channel layouts available in MPEG-H smaller than the target layout.

Before exporting audio and metadata to the MPEG-H production format, labeling presets and components in different languages provides good service for the multilingual dissemination of audiovisual content. Depending on the preferred language settings on their target device, users will see the labeling in their respective language accordingly.

### 4.3. Suggestions for Interactivity in Classical Music

We implemented two possible user interactivity scenarios in this production in order to show possibilities and encourage content producers to invent further scenarios for recipients:

- (1) Interactive choice of seats
- (2) Multilingual music commentary

For the interactive choice of seats, users could choose between presets called *Conductor* for a close-up orchestral sound, *Queen's seat* for the default mix and *Balcony* for a predominantly spatial experience enhancing the diffuse, enveloping sound as can be seen in Fig. 6. To implement this, we split the mixed signals in four components, separating the basic mix from diffuse surround and height signals:

1. Two mono objects panned to  $\pm 30^\circ$  comprising the front left and right signal respectively (front main system, spot microphones, reverberation).
2. A 5-channel object comprising the front center signal (front main system, spot microphones), surround reverberation and the height layer front signal
3. A 2-channel object panned to  $\pm 110^\circ$  comprising the mid layer diffuse surround microphones
4. A 4-channel object panned to  $\pm 30^\circ$  and  $\pm 110^\circ$  with elevation at  $35^\circ$  (height speaker positions) comprising the height layer surround microphones, Hamasaki square and the 4 channel upper layer reverberation.

Changing the relative level between the components conveys the impression of sitting closer or further away from the orchestra. When choosing the *Conductor* preset, left and right objects from the first component are spread in panorama using dynamic position metadata. This creates the impression of sitting in the middle of the orchestra.

As a second interactive use case, we produced a music commentary that users can switch on and off, choosing between two different languages. The commentators shed light on compositional and content-related aspects of the pieces. Theoretical backgrounds of the compositions are thus made comprehensible to a broad audience. The commentary offers an additional informative level that combines entertainment and cultural education. Users are further allowed to change volume and position within restrictions that we set in the MPEG-H Authoring Plugin. For the multilingual music commentary, we created a 4+5+0 channel bed

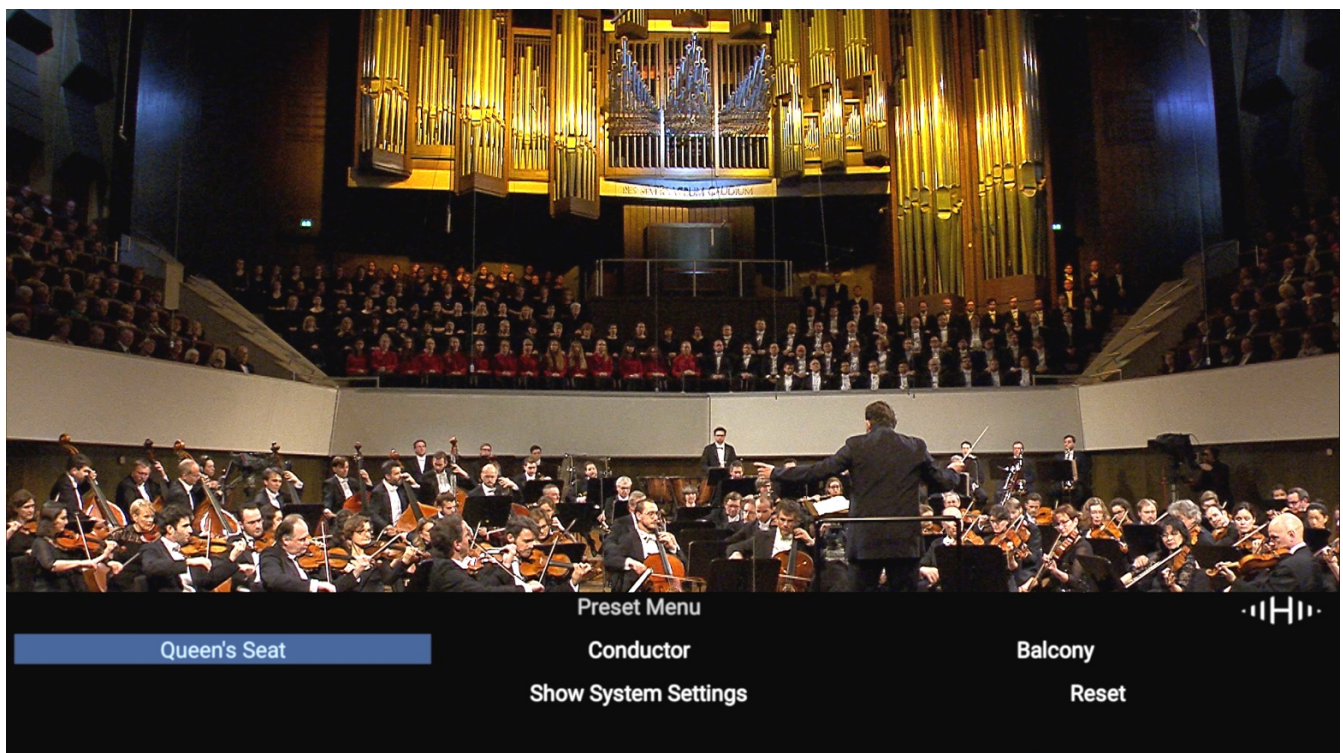


Fig. 6: On-screen display on the user's TV set to show available interactivity options. The top line (Queen's Seat, Conductor, Balcony) represents pre-configured mix presets to choose from.

comprising the music mix (as described in 4.2) and two mono objects containing the commentary. User interaction restraints were set, so volume and position can be modified within certain limitations: We allowed commentary gain shifts between -2 dB and +4 dB and position changes in the horizontal layer of  $\pm 110^\circ$  and between  $0^\circ$  and  $+35^\circ$  in the vertical layer. The default position was set to  $0^\circ$  in both, vertical and horizontal layer equivalent to the M0 or center speaker in a 4+5+0 speaker layout.

To ensure speech intelligibility, the channel bed is lowered by the decoder, as soon as commentary signal is present, comparable to a normal voice-over mix. This is attained by gain metadata, so called *gain sequences*, that are attached to the regarding preset and are applied in the decoder to a pre-assigned audio component. Gain sequences are comparable to volume automation. They are encoded and transported within the regular MPEG-H stream. This allows for sending only one, unattenuated bed mix and the related objects. Depending on the user preset selected, the respective objects are activated and their gain sequences process the mix accordingly.

## 5 Conclusions

This paper describes our field test for immersive and interactive audio production in classical music at Gewandhaus zu Leipzig. We recorded three different concerts, i.a. Beethoven's 9<sup>th</sup> symphony conducted by Gewandhauskapellmeister Andris Nelsons. We found convincing microphone setups for the immersive recording of classical symphonic music in Gewandhaus. Elevated sources like choir and soloists were captured by our height microphone system and supported by their corresponding spots in the mix. Elevation and spatial fidelity were rated credibly authentic in informal expert listening evaluations. As the main microphone system, a combination of Decca Tree and a more diffuse spaced AB was used for the mid layer. Broad cardioids were used in a spaced AB system for the front height channels. Surround microphones were widely spaced with omnidirectional microphones as mid-layer surrounds and broad cardioids in the height layer. We used a Hamasaki square as room microphone system for the height layer.

Beyond that, we specified the common process for mastering and authoring classical music using MPEG-H. Downmix compatibility as well as loudness consistency are ensured by MPEG-H active downmix and loudness adaption algorithms respectively. Challenges can occur when exporting individual items with differing loudness and compiling them to one program. We suggested a solution using *album loudness* and detailed the further procedure up until the encoding process, using MPEG-H specific software.

Furthermore, we suggested two possible scenarios for user-interactivity: Choice of seats within the concert hall and a commentary which explains the theoretical background to the composition within an edutainment context. From the concerts recorded, we created AV- and audio-only demos, which show that MPEG-H is suitable for the presupposed requirements.

We recommend to put further work into the microphone setup research and optimization. Also, NGA music mastering and authoring processes should be further investigated and optimized. The evaluation of the described interactivity scenarios is another topic to investigate further.

## 6 References

- [1] ISO/IEC 23008-3, *High efficiency coding and media delivery in heterogeneous environments — Part 3: 3D Audio*, Geneva, Switzerland: International Standard, 2019.
- [2] J. Herre, J. Hilpert, A. Kuntz et al. "MPEG-H Audio - The New Standard for Universal Spatial / 3D Audio Coding," in *137th AES Convention*, Los Angeles, USA, 2014.
- [3] G. Theile, "Multichannel Natural Music Recording Based on Psychoacoustic Principles," Bolkesjø, Norway, 2001.
- [4] F. Camerer and C. Sodl, "Classical Music in Radio and TV - a multichannel challenge," *The ORF Surround Listening Test 2001*, 2001.
- [5] K. Hamasaki, K. Hiyama and R. Okumura, "The 22.2 Multichannel Sound System and Its Application," in *118th AES Convention*, Barcelona, Spain, 2005.
- [6] K. Hamasaki and W. V. Baelen, "Natural Sound Recording of an Orchestra with Three-dimensional Sound," in *138th AES Convention*, Warsaw, Poland, 2015.
- [7] A. Gernemann-Paulsen, "Decca-Tree - Gestern und Heute (The Decca-Tree - Past and Present)," [www.uni-koeln.de/phil-fak/muwi/ag/tec/deccatree.pdf](http://www.uni-koeln.de/phil-fak/muwi/ag/tec/deccatree.pdf), Universität Köln, 2002.
- [8] S. Weinzierl, *Handbuch der Audiotechnik*, Berlin - Heidelberg: Springer Verlag, 2008.
- [9] F. Rumsey, S. Zielinski and R. Kassier, "On the relative importance of spatial and timbral fidelities in judgements of degraded multichannel audio quality," *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. p. 968-976, 2005.
- [10] R. Sazdov, "The effect of elevated loudspeakers on the perception of engulfment and the effect of horizontal loudspeakers on the perception of envelopment," in *International Conference on Spatial Audio (ICSA)*, Detmold, Germany, 2011.
- [11] D. Griesinger, "General Overview of Spatial Impression, Envelopment, Localisation and Externalisation," in *AES 15th International Conference*, Copenhagen, Denmark, 1998.
- [12] G. Theile and H. Wittek, "Die dritte Dimension für Lautsprecher-Stereofonie (The Third Dimension for

- Loudspeaker Stereophony)," *VDT Magazin*, pp. p. 31-37, 02 2011.
- [13] U. Scuda, "Comparison of Main Microphone Systems for 3D-Audio Recording," in *28th Tonmeistertagung - VDT International Convention*, Cologne, Germany, 2014.
- [14] Y. Grewe and U. Scuda, "Comparison of Main Microphone Systems for 3D-Audio Recordings," in *29th Tonmeistertagung - VDT International Convention*, Cologne, Germany, 2016.
- [15] T. Nishiguchi, R. Okumura and Y. Nakayama, "Production and Live Transmission of 22.2 Multichannel Sound with Ultrahigh-definition TV," in *122nd AES Convention*, Vienna, Austria, 2007.
- [16] H. Stenzel and U. Scuda, "Producing Interactive Immersive Sound for MPEG-H: A Field Test for Sports Broadcasting," in *137th AES Convention*, Los Angeles, USA, 2014.
- [17] A. Silzle, "The EU Project ORPHEUS: Object-based Broadcasting - for Next Generation Audio Experiences," *VDT Magazin*, vol. 1, no. ISSN: 2509-5927, pp. 24-27, 2017.
- [18] EBU-TR-042, "Example of an End-to-end OBA Broadcast Architecture and Workflow," European Broadcasting Union, Geneva, Switzerland, 2018.
- [19] A. Silzle, R. Schmidt, W. Bleisteiner et al. "Quality of Experience Tests of an Object-Based Radio Reproduction App on a Mobile Device," *J. Audio Eng. Soc.*, vol. 67, no. 7/8, pp. 568-583, 2019.
- [20] C. Simon, Y. Grewe, N. Faecks et al. "Field Tests for Immersive and Interactive Broadcast Audio Production using MPEG-H.," *Set International Journal of Broadcast Engineering*, vol. 4, pp. 40-46, 2018.
- [21] "EBU Recommendation R 128: Loudness Normalisation and Permitted Maximum Level of Audio Signals," European Broadcasting Union, Geneva, Switzerland, 2014.
- [22] "EBU Tech 3343 v3, Guidelines for Production of Programmes in Accordance with EBU R 128," European Broadcast Union, Geneva, Switzerland, 2016.
- [23] "ITU-R Recommendation BS.1770-4, Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level," Intern. Telecom Union, Geneva, Switzerland, 2015.
- [24] "ATSC Doc. A/85, ATSC Recommended Practice: Technique for Establishing and Maintaining Audio Loudness for Digital Television," 2013.
- [25] ITU-BS-2076, "Audio Definition Model," in *ITU Recommendation*, Geneva, Switzerland, 2017.
- [26] R. Bleidt et al. "Development of the MPEG-H TV Audio System for ATSC 3.0," *IEEE Transactions on broadcasting*, vol. Vol. 63, no. No. 1, pp. 202 - 236, 2017.
- [27] A. Silzle, S. Geyersberger et al. "Vision and Technique Behind the New Studios and Listening Rooms of the Fraunhofer IIS Audio Laboratory," in *126th AES Convention*, Munich, Germany, 2009.
- [28] D. B. Levy, *Beethoven: The Ninth Symphony (Revised Edition)*, New Haven and London: Yale University Press, 2005.
- [29] W. Howie and R. King, "Exploratory Microphone Techniques for Three-dimensional Classical Music Recording," in *138th AES Convention*, Warsaw, 2015.
- [30] H. Lee and C. Gribben, "Effect of Vertical Microphone Layer Spacing for a 3D Microphone Array," *J. Audio Eng. Soc.*, vol. 62, no. 12, pp. 870-848, 2014.
- [31] M. Rose, "MPEG-H Authoring Plug-in 2.0," Fraunhofer IIS, 2019. [Online]. Available: <https://www.iis.fraunhofer.de/de/ff/amm/dl/software/mhapi.html>. [Accessed 29 07 2019].
- [32] ISO/IEC 23003-4, *MPEG Audio Technologies Part 4: Dynamic Range Control*, Geneva, Switzerland, 2015.
- [33] W. Howie, R. King and M. Boerum, "Listener Preference for Height Channel Microphone Polar Patterns in Threedimensional Recording," in *139th AES Convention*, New York, USA, 2015.
- [34] N. Zacharov, C. Pike and F. Melchior, "Next Generation Audio System Assessment Using the Multiple Stimulus Ideal Profile Method," in *8th International Conference on Quality of Multimedia Experience (QoMEX)*, Lisbon, Portugal, 2016.
- [35] T. H. Pedersen, and N. Zacharov, "The Development of a Sound Wheel for Reproduced Sound," in *138th AES Convention*, Warsaw, Poland, 2015.
- [36] F. Kuech, M. Kratschmer, B. Neugebauer et al. "Dynamic Range and Loudness Control in MPEG-H 3D Audio," in *139th AES Convention*, New York, USA, 2015.
- [37] Y. Grewe, C. Simon and U. Scuda, "Producing Next Generation Audio using the MPEG-H TV Audio System," in *NAB*, Las Vegas, 2018.
- [38] ISO/IEC/JTC1/SC29/WG11, "MPEG2014/N14462 Active Downmix Control," València, Spain, 2014.





## Abstract Reviewed Paper at ICSA 2019

Presented\* by VDT.

### Capturing 3D Audio: A pilot study on the spatial and timbral auditory perception of 3D recordings using main-array and front-rear separation in diffuse field conditions

J. Grab<sup>1</sup>

<sup>1</sup> *SAE Institute London, United Kingdom, Email: janetmariagrab@gmail.com*

#### Abstract

This research is a preliminary pilot experiment into the subjectively perceived differences between the recordings resulting from three different 3D microphone arrays: A Bowles-Array (Main-Array) with a vertical coincident height-channel microphone layer, a Fukada-Tree/Hamasaki-Cube configuration (F/R-Array) and a Hybrid-Array containing the signals from the Bowles-Array main layer and the Hamasaki-Cube height layer. It was hypothesised that the arrays in concern will produce recordings that shall lead each to an increased perception of specific attributes for all sources tested (cello, violin, handpan, djembe, guitar). In order to detect possible patterns in spatial and timbral auditory perception subjective listening tests included direct scale magnitude estimations for the attributes Naturalness, Presence, Preference, Width, Localisation Accuracy, Distance/Depth, Envelopment, Spatial Balance, Room Perception, Vertical Image Shift, Vertical Image Spread, and Vertical Frequency Separation and category scaling for the assessment of timbral attributes. Results suggest that none of the arrays in concern conveyed an increased perception of any of the attributes for all sources, which disproves the hypothesis and indicates a source-dependent performance. Simultaneously patterns in the subject responses have been detected which could be explained through psychoacoustic findings focussing on the correlation of perception between the attributes in question. Furthermore, by trying to explain the obtained differences in auditory perception between the different arrays, some assumptions could be made upon what components of which array could have contributed to a specific perception. These findings could serve as a reference for future experiments in the fields of 3D recording techniques or psychoacoustics.

#### 1. Introduction

The current investigation entails a comparison between a main array technique being a Bowles-Array (Main-Array) and a system with front-rear separation being a Hamasaki-Cube/Fukada-Tree configuration (F/R-Array). For the sake of experimentation, a hybrid version containing the signals of the Bowles-Array for the main layer and the signals of the Hamasaki-Cube for the height layer was included. The selection of these arrays was based on previous research [1-15], which at the same time was the foundation of certain assumptions on their behaviour for specific attributes:

The Bowles-Array was expected to convey an enhanced perception of Naturalness, Presence and Vertical Image Spread (VIS), a preference regarding Timbre, an enhanced

risk of a Vertical Phantom Image Shift, and a reduced perception of Spatial Balance and Distance/Depth. The Fukada-Tree/Hamasaki-Cube, on the other hand, was thought to favour an enhanced perception of Envelopment, Width, Spatial Balance and Room Perception, a reduced possibility of a VIS, a reduced risk of a Vertical Phantom Image Shift, and a more stable Localisation Accuracy. In addition, the Hybrid-Array was assumed to show a reduced risk of a Vertical Phantom Image Shift, a reduced possibility of a VIS, and a limited perception of Distance/Depth. Besides, there was a possibility of tonal colouration for all three arrays although the nature of these colourations was unknown by the time to the authors' knowledge [10, 12, 14].

## 2. Methodology

### 2.1. Preliminary Recording Session

Prior to the recording session, a test session was organised in Bankstock Studios to ensure that the chosen angle between the main and height layer of the Main-Array would achieve an ICLD (interchannel level difference) of at least -9.5dB. This was found to be the localisation threshold for an ICTD (interchannel time difference) of 0ms and thus corresponds to the vertical coincident microphone placement of the setup in question [14]. Although the acoustic properties of the recording space are different, the acoustically controlled environment would at least allow for indications whether such an ICLD could be achieved with the desired microphone angles and directivities. The result of the test session suggested that an ICLD of -9.5dB would be realistic.

### 2.2. Recording Procedure

The recording took place at All Saints East Finchley, London. For the recording, all microphone signals were routed to three RME Octa Mic II preamps, an Antelope Orion 32 interface, and recorded into Pro Tools in PCM wave format at 96kHz/24bit resolution. To allow for consistency throughout the comparison of the different techniques a modular system was the preferred choice as it allows for different polar patterns while maintaining the same preamps (Schoeps Colette Series, CMC6 preamps, capsules, KA40 diffraction attachments). As having used the same microphone model for all channels, matching the microphone sensitivity was considered achieved apart from slight sensitivity differences caused by the different capsules. The polar patterns, however, were considered an experimental constant. During the recording, the goal was to match the input gain on all channels to minimise the differences of possible colouration amongst the channels and to maintain the natural level relationships between the channels. However, when strictly adhering to this the SNR (signal-to-noise ratio) of the Hamasaki-Cube would exceed acceptable limits. Therefore, the gain applied to the channels of the Hamasaki-Cube was matched within its low and height layer but was higher than the gain of other channels.

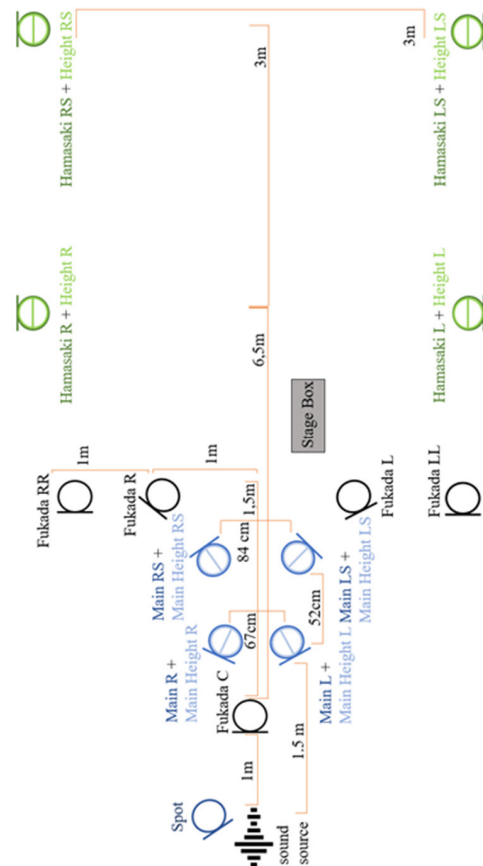


Fig. 1: Floor plan, top view.



Fig. 2: Recording setup, top view.

Position	Capsule on Schoeps CMC 6	Axis	Height
Main L	MK2H	30° outwards, 45° to source	2m
Main R	MK2H	30° outwards, 45° to source	2m
Main LS	MK2H with KA40	30° outwards, 45° to floor	2m
Main RS	MK2H with KA40	30° outwards, 45° to floor	2m
Main Height L	MK4	110° from main upwards	2m
Main Height R	MK4	110° from main upwards	2m
Main Height LS	MK4	110° from main upwards	2m
Main Height RS	MK4	110° from main upwards	2m
Fukada L	MK4	45° outwards	1.8m
Fukada C	MK4	0°	1.8m
Fukada R	MK4	45° outwards	1.8m
Fukada LL	MK2	0°	1.8m
Fukada RR	MK2	0°	1.8m
Hamasaki L	MK8	Positive lobe 90° outwards	3m
Hamasaki R	MK8	Positive lobe 90° outwards	3m
Hamasaki LS	MK8	Positive lobe 90° outwards	3m
Hamasaki RS	MK8	Positive lobe 90° outwards	3m
Hamasaki Height L	MK41	0° to ceiling	5m
Hamasaki Height R	MK41	0° to ceiling	5m
Hamasaki Height LS	MK41	0° to ceiling	5m
Hamasaki Height RS	MK41	0° to ceiling	5m
Spot	MK4	source-dependent	-

Tab. 1: List of microphones.



Fig. 3: Recording setup, side view.

### 2.3. Listening Test Design

The attributes have been selected and defined based on [8, 12, 16-20]. The chosen response format for the evaluation of the spatial attributes was a direct scale magnitude estimation where the subject assigns a (numerical) value to one stimulus and then judges subsequent stimuli against the first. Timbral attributes, however, have been graded with category scaling, where the subject is asked to assign a category (in this case a timbral or dynamic label) to each stimulus presented.

Double-blind multiple stimuli comparison tests were conducted using a GUI with the Huddersfield Universal Listening Test Interface Generator [21]. The subject could freely turn its head if it stayed in the sweet spot. The task was to grade three stimuli against each other on a continuous rating scale. The scale ranged from 0 (labelled “lesser”) to 100 (labelled “greater”), whereas the stimulus with the “greatest” attribute impression was taken as a reference of 100 with the other two stimuli being graded accordingly. This procedure was proposed in [22] to reduce scaling bias. The presentation order of both, the stimuli and the trials was randomised to avoid potential biases. The stimuli were synced in playback, meaning the subject could switch between mixes at any point during the playback. For each test, the subject was to complete a total of five trials, each of which contained the stimuli of the different mixes. When rating Preference, the subject was not given any specific subjective qualities or attributes to consider but was advised to write a comment on its decision on a paper.

### 2.4. Reproduction Setup

All audio playback including mixing took place in the Auro-3D studio at SAE Brussels, an acoustically treated listening room with a 10.1 Auro-3D setup with Sonodyne SM100AK speakers. The sound pressure level of each loudspeaker was measured and calibrated for 79dB SPL at the listening position using pink noise. The monitor level was calibrated and kept constant to -18dBFS throughout the sessions.

### 2.5. Stimuli Creation

Musical excerpts of 30 seconds were chosen. The selection was mostly aimed towards passages containing pauses of a certain length to ease the perception of room-related attributes. Based on the concept of “ecological validity” [19], a balanced mix was chosen for stimuli creation. Phase relations have been checked aurally and visually, and the polarity was flipped in a few cases where necessary. The stimuli have been mixed in Pro Tools with the Auro-3D Authoring Tools through the Digidesign D-Control (ICON) control surface. Due to the nature of the research questions, no processing was applied. The only exception was a LPF at 250 Hz on the Fukada LL and RR channels, as proposed by Rumsey [23]. All stimuli have been level matched for ± 0.2dB.

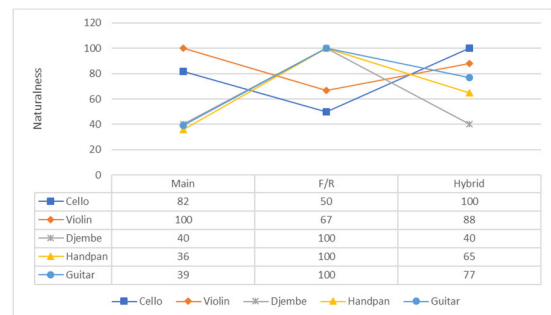
### 2.6. Limitations

Due to limited resources, the process of subject pre- and post-selection could not be accomplished to obtain a listening panel of at least five expert subjects to ensure a sufficient resolution in the test [24]. However, a preliminary listening test was

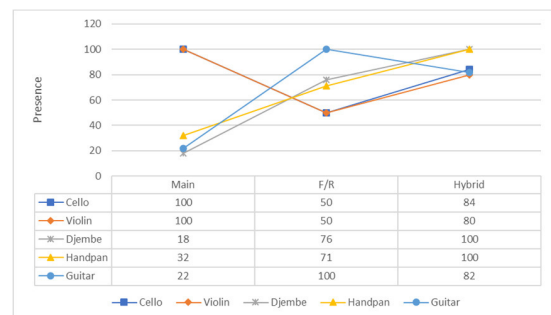
conducted one day before the listening test, and the results of the two tests showed a strong correlation, which indicates intra-subject consistency. Similarly, the validity of the results can be seen reduced as only one balancing engineer was involved, and not several, to minimise the factor of subjectivity in stimuli creation. Also, as the subject was at the same time the experimenter, the subject was familiar with the experimental detail and thus more prone to expectation bias. However, the results throughout the listening tests for the different array-mixes suggest that the magnitude of expectation bias was not dominating the subject’s evaluation. Furthermore, the recording was conducted in one space only. Concerning that, the ecological validity of the experiment would be improved when involving ensembles. In addition, based on the recording setup, the results are only applicable to a dry-wet scenario, and not a sound all-around approach.

## 3. Results

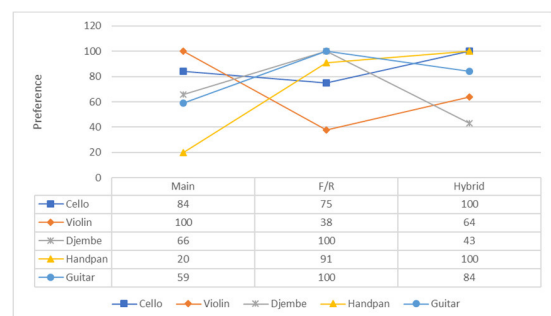
### 3.1. Graphical Representations of the Listening Test Results



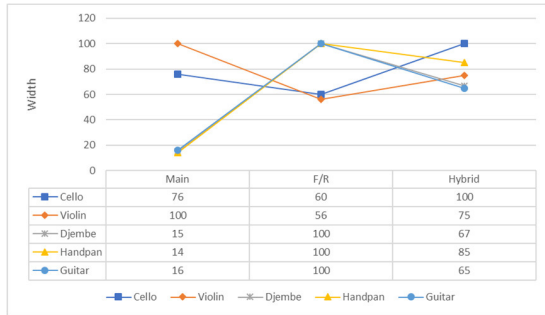
Tab. 2: Results for the attribute Naturalness.



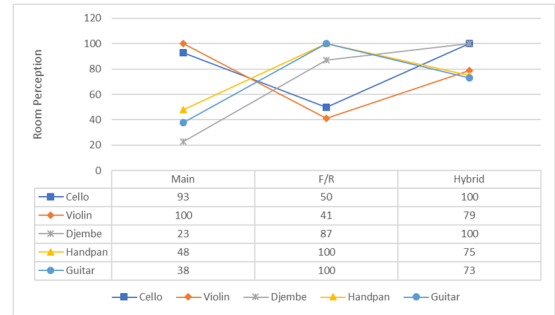
Tab. 3: Results for the attribute Presence.



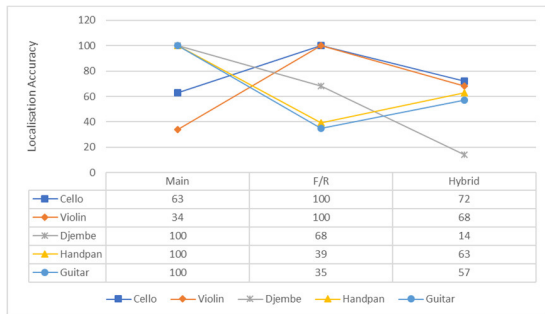
Tab. 4: Results for the attribute Preference.



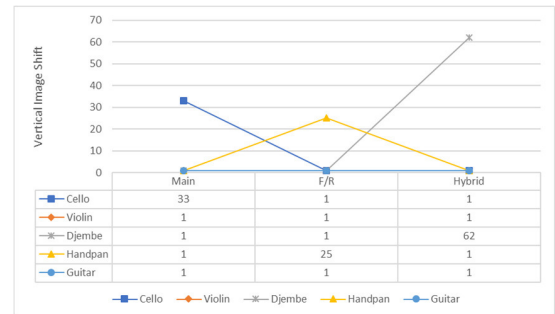
Tab. 5: Results for the attribute Width.



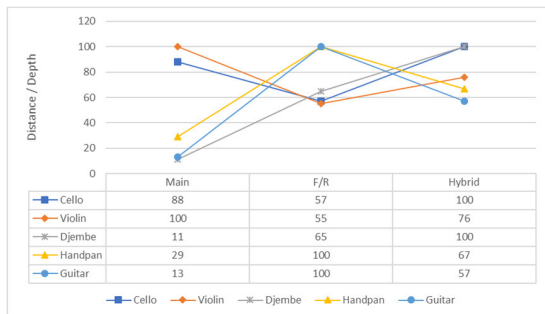
Tab. 10: Results for the attribute Room Perception.



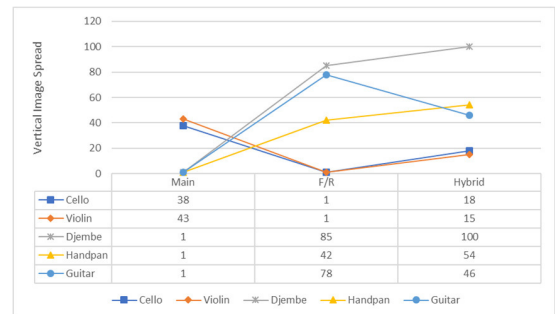
Tab. 6: Results for the attribute Localisation Accuracy.



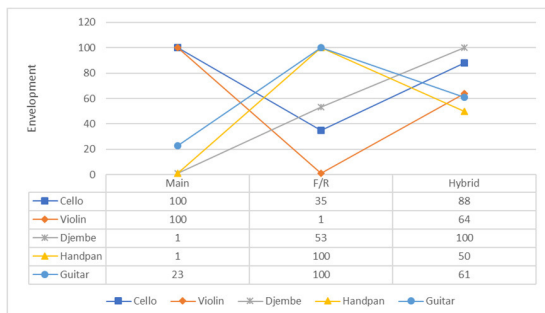
Tab. 11: Results for the attribute Vertical Image Shift.



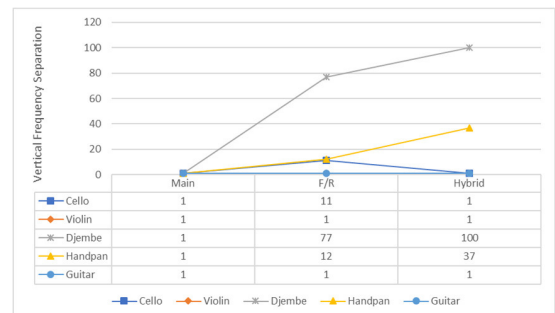
Tab. 7: Results for the attribute Distance/Depth.



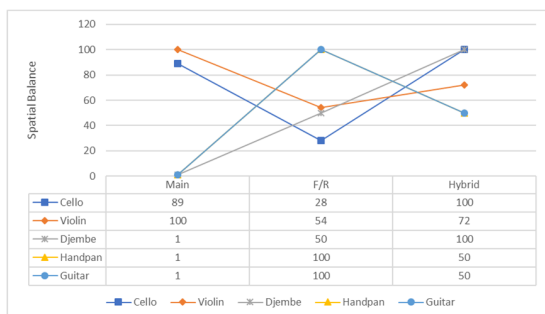
Tab. 12: Results for the attribute Vertical Image Spread.



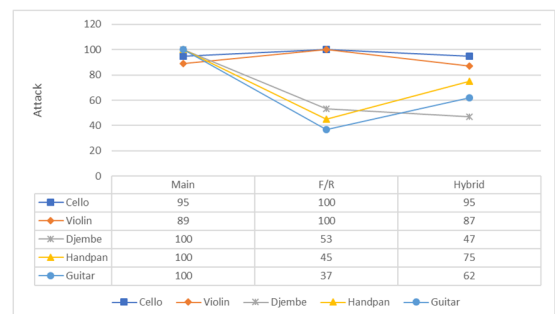
Tab. 8: Results for the attribute Envelopment.



Tab. 13: Results for the attribute Vertical Frequency Separation.



Tab. 9: Results for the attribute Spatial Balance.



Tab. 14: Results for the attribute Attack.



Source	Stimulus used for scale A	Stimulus used for scale B	Stimulus used for scale C
Handpan	thin	full, but lacking mids	slightly nasal, less full than B, but more homogeneous, natural, highest treble content in reverb
Cello	full	homogeneous, natural, highest treble content in reverb	(very) thin, nasal
Djembe	homogeneous, natural, highest treble content in reverb	full, but completely lacking mids, excessive bass	canny/nasal, thin
Violin	treble strength neutral	sharp, thin, nasal	brilliance, highest treble content in reverb
Guitar	lacking mids, nasal, thin	homogeneous, brilliant, full	completely lacking midrange, very bright

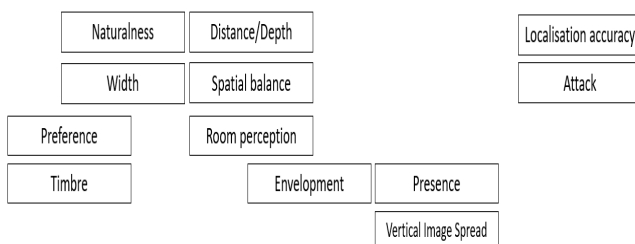
**Tab. 15:** Results for the timbral attribute category scaling. The colour of blue indicates the mix of the Main-Array, orange the mix of the F/R-Array, and green the mix of the Hybrid-Array.

Source	Stimulus used for scale A	Stimulus used for scale B	Stimulus used for scale C
Guitar	<b>very open, yet still clear, wide pleasing sound</b>	very precise but lacking space	compromise between A and B
Violin	<b>very open sound, good D/R ratio, not sharp</b>	not as sharp as C but not as wide and opened as A	sharp, edgy, lacking space
Djembe	phase issues not as bad as B, but also not as realistic as C	disturbing frequency, phase issues in the ambient sound with changing frequency	<b>very realistic</b>
Cello	dislike timbre	<b>balanced D/R ratio, further away</b>	not as wide as B, lows too present, narrowing the picture <b>unaturally natural, real</b>
Handpan	natural but less than C	no space	

**Tab. 16:** Comments on the grading for the attribute Preference. Comments in bold indicate the comment of the preferred stimulus.

### 3.2. Correlation between Attributes and Subject Responses

Figure 4 depicts a schematic representation of the correlations found between the assessed attributes based on their response patterns through direct scale magnitude estimation (what array was graded how for what source regarding a specific attribute). Vertically coherent displayed attributes showed an identical response pattern regarding the ranking of the different arrays. Horizontally intended attributes showed an identical response pattern to the attributes above, except for one source. Similarly, the closer the attributes are horizontally to each other, the more similar were their response patterns. The attributes Vertical Frequency Separation and Vertical Image Shift have not been included in this graphic as the obtained values do not allow for a direct comparison with the other attributes.



**Fig. 4:** Correlations between attributes based on their response patterns.

### 3.3. Correlations between Array Gradings and Sources

First, the Main-Array dominated the high gradings for the violin regarding all attributes. Exceptions are Localisation Accuracy and Attack where the Main- and F/R-Array seemed to have swapped their behaviour. The Main-Array also prevailed the highest ratings for the violin and cello regarding the attributes Envelopment, Presence and Vertical Image Spread. Similarly, the F/R-Array dominated responses of high

values for the guitar except for Localisation Accuracy and Attack. Almost to the same extent the F/R-Array led to high scores for the handpan. Analog to these examples, the Hybrid-Array was responsible for most high ratings for the cello and djembe. Considering the sonic nature of the instruments when describing source-array dependencies, the following regularities have been discovered: The Main-Array seemed to dominate primarily the responses of the violin, which exhibits a sustained HF character. Besides, it also featured sustained sources with different frequency content (violin/cello). Analog, the F/R-Array seemed to prevail sources mostly being active in the mid-frequency range and being of both, a sustained and percussive nature. The Hybrid-Array, on the other hand, seemed to have the most influence on high ratings of instruments with enhanced LF content. Also, the Main-Array led to responses of lower scores mostly for instruments entailing a percussive element. An interesting note is also the low values for the Localisation Accuracy of the violin, as for all other attributes the Main-Array resulted in the highest ratings. Secondly, a minor tendency of the F/R-Array towards lower gradings on string instruments can be assumed. Regarding the attributes Vertical Image Shift and Vertical Frequency Separation the only pattern to be discovered was that these phenomena appeared almost exclusively in the F/R- and Hybrid-Array, which both contain the height layer of the Hamasaki-Cube.

## 4. Discussion

### 4.1. Correlation between Attributes and Subject Responses

The relationships between attributes as depicted in Figure 4 are confirmed by previous research and will be discussed hereafter. Since the attribute clusters in Figure 4 are based on the rankings of the arrays for the different sources, a close correlation between the attributes implicates an identical or strongly correlated response pattern for the different arrays. Hence, when the attribute correlations can be backed up by previous research, the validity of the array gradings can be seen increased. Naturalness, for example, was proven in experiments “to be by far the most important factor in determining overall preference in sound quality [...] and it may have a strong timbral component and be highly context dependent” [23]. Thus, this statement confirms the close relation between Naturalness, Preference and Timbre. It also explains the frequent appearance of the descriptor “natural” in the comments on Preference. As it was shown that timbral fidelity contributes strongly more to the overall quality judgment than spatial fidelity [23, 25], the identical response pattern for Preference and positive timbral descriptors leading to a higher correlation between Preference and Timbre than Preference and any spatial descriptive attributes can be explained this way. The positive timbral descriptors used when describing the auditory perception for stimuli which also had been graded highest in Preference were “homogeneous”, “natural”, “brilliant” and “highest treble content in reverb.” When considering that Naturalness is one of the most critical factors for Preference and has been defined as “how similar to a natural listening experience the sound as

a whole sounds,” a possible explanation for the negative correlation between negative timbral perception and Preference could be seen in anomalies being specifically related to reproduced sound, such as phase issues, e.g. These are rarely experienced in natural environments and thus lack a reference point against which to compare these experiences [23]. Unlike Timbre, Naturalness appeared mostly as a positive descriptor in the comments for Preference and could be explained through the same concept [19]. Furthermore, the perception of space as a further influential factor for Preference was confirmed by Toole who found that increased quality of spatial ratings can significantly influence the overall sound quality rating [26]. In that regard, the descriptor “open”, which frequently appears in the comments on Preference, contributes to “a feeling of space” and leads to higher ratings for Naturalness and Preference [23]. Hence, the rather close relation between the spatial descriptive attributes and the cluster of Naturalness and Preference can be seen confirmed. Having backed up the strong correlation between the array ratings of Preference, Naturalness and Timbre, their correlation to the attribute Width must be included in the discussion. In concert hall acoustics ASW (apparent source width) has been associated with positive listener responses [19]. According to the definitions of the current study ASW could be considered equal to “individual source width”, which is a sub-attribute of Width. Therefore, the concept of ASW could be seen to back up the correlation between Width and Preference. In relation to that, the stronger correlation between Width and Preference compared to Depth and Preference was previously confirmed in [19]. Although environmental depth and source distance seem to be dominated by the perception of environmental width, Depth is still crucial to the appreciation of sound quality [23]. This explains why the array gradings for Depth are not as correlated to the array gradings of Preference as are the gradings for Width but can still be found in the same cluster area in Figure 4. Another phenomenon to explain is the correlation between the spatial descriptive attributes, Envelopment and Presence. According to Rumsey, the link of these attributes to Preference, Naturalness and so on can be established in the connection between Width and Room Perception. From there, on one side, he found a correlation between Envelopment, Room Perception and Spatial Balance, and on the other side, he declares that “presence and environmental envelopment are not necessarily the same, although they may be closely related” [19]. Combining these statements, it can be argued that the function of Envelopment as a link between spatial descriptive attributes and Presence in Figure 4 can be seen confirmed. Last but not least the isolated cluster of Localisation Accuracy and Attack requires further explanation. Presumably, a “clear transient response” as by definition for Attack in the listening test would lead to a better Localisation Accuracy. This might explain why these attributes achieved an identical response pattern. The outcome that their response patterns seem uncorrelated to the response patterns of all other attributes was confirmed in the findings of Berg & Rumsey having proved “that localisation in itself is not the attribute closest to naturalness and positive sensations” [27]. Therefore, by having backed up the correlations found

between the attributes as depicted in Figure 4 the nature of the response patterns for the different arrays could be considered validated.

## 4.2. Expected and Unexpected Results

Although the patterns of the listening test were found to be consistent, confirmed by research, and intra-subject consistency is assumed there is a rather distinctive deviation from the expectations of the different arrays outlined in the introduction.

### 4.2.1. Source-Dependent Array Behaviour

It was assumed that each array would dominate the high ratings for all sources for specific attributes. Therefore, the outcome of the listening test indicating that no array dominates the highest scores for all sources for a specific attribute (or at least four of the five sources) and therefore giving the results a source-dependent character, was not expected. Although some regularities regarding frequency content and acoustic envelopes could be identified amongst the sources, a thorough explanation of the source-dependent results requires further experiments with a more controlled experimental design. When investigating the source-dependence in the psychoacoustic realm, research indicates that the radiation pattern of the different sources could be a factor which could have influenced the current results. Martin *et al.* proved that the radiation pattern impacts the instrument’s perceived audio image whereas the non-coincident arrays featured the most irregular source image perceptions [28]. This is worth mentioning as in the current study only non-coincident arrays have been applied. Although the research of Martin *et al.* was only concerned about imaging, the diverse perception of the source images within the same arrays indicates that also other attributes could be affected by radiation patterns. Even if the approach of radiation patterns won’t lead to an explanation of the current results, the insight gained therein could provide a better understanding of 3D recording, according to Bowles [1].

### 4.2.2. Vertical Image Spread, Vertical Image Shift and Vertical Frequency Separation

The unexpected result of the F/R- and Hybrid-Array dominating the perception of Vertical Image Spread, Vertical Image Shift and Vertical Frequency Separation, whereas all these attributes have been initially assigned exclusively to the Main-Array, could be explained when having a look at psychoacoustics. Spectral graphs obtained of the Hamasaki-Cube signals indicate a slightly enhanced HF-content compared to the ambient signals of the Main-Array. As the F/R- and Hybrid-Array both contained the signals of the Hamasaki-Cube for their height layer, this could have caused a pitch-height effect [10], which, amongst other possible factors, may have led to the perception of VIS, Vertical Image Shift or Frequency Separation. The scenario of an exceeded localisation threshold during mixing leading to vertical ICCT and thus to these effects [10, 14] can be considered unlikely, as the Hamasaki-Cube is optimised to capture mainly ambient sound [3]. A remaining question here would be why some of these effects have been observed in the Main-Array for string

instruments, but no other sources. As the ICCT leading to these effects depends on the ICLDs between the main and height layer signals [14], this would indicate the option that the ICLDs of the string instruments were smaller than the ICLDs of the other sources. However, the reason for this would be currently unknown to the author.

#### 4.2.3. Naturalness and Width

Although the Main-Array was claimed to convey an enhanced perception of Naturalness, in fact, only one source scored highest for this attribute for the Main-Array. This implies that the F/R- and Hybrid-Array dominated the attribute Naturalness. A possible reason could be that the highly decorrelated signals of the Hamasaki-Cube, being a part of both, the F/R- and Hybrid-Array, lead to a decreased IAC (interaural cross correlation) and thus to an increased ASW [2]. ASW can be considered being part of the attribute Width as per definition used, which resulted in the same response pattern as Naturalness. Having said this, the expected result of the F/R-Array achieving the highest scores for Width may be explained the same way.

#### 4.2.4. Room Perception, Spatial Balance and Distance/Depth

Although it was the Hybrid-Array scoring highest for Room Perception, Spatial Balance and Distance/Depth, and not the F/R-Array, as expected, it can be said that the Main-Array scored considerably lower than any of the other two arrays for these attributes, even if this was expected. It could be assumed, that since both, the F/R- and the Hybrid-Array share the common ground of the Hamasaki-Cube height layer, that a possible explanation could be the way early reflections have been captured, as stated in [3] and [13]. Therefore, one could think that the increased height, distance and directivity of the Hamasaki-Cube microphones compared to the directivity and placement of the Main-Array height layer microphones would have led to a more distinctive capture of early reflection key parameters, such as mentioned in [13]. As a consequence, this would ease Room Perception and the perception of Spatial Balance and Distance/Depth. This, however, does not explain why the Hybrid-Array entailing omnidirectional microphones in the main layer and thus having no directivity in its capture could score higher than the F/R-Array containing the Hamasaki-Cube main layer which is optimised for capturing lateral early reflections. This is an unexpected outcome as the increased importance of lateral early reflections compared to ceiling reflections for Room Perception was originally found by Barron [29], although his experimental design slightly differs from the current study.

#### 4.2.5. Envelopment

Based on the same arguments no explanation could be found why the Hybrid-Array scored higher on average than the F/R-Array for Envelopment. Although it is not early lateral reflections influencing the perception of Envelopment, but late lateral reflected energy [13, 19], the parameter of directional lateral capture remains the same. What could be explained instead is the positive correlation between the average scores of Envelopment and Spatial Balance, based on

Hanyu *et al.* [4]. Consequently, as the Hybrid-Array was perceived to have the highest degree of Spatial Balance, it thus might also have been perceived as most enveloping.

#### 4.2.6. Presence

Similarly, based on [19], the increased perception of Spatial Balance of the Hybrid-Array might have led to an increased perception of Presence compared to the other arrays. Based on the rather high ratings for Presence of both, the Hybrid- and F/R-Array compared to the Main-Array, it can only be hypothesised that the increased treble content in the Hamasaki-Cube height layer might have contributed to the perception of “realism” and thus to Presence [18]. If this would be the case, however, the question would arise why the Hybrid-Array was scored higher than the F/R-Array as the F/R-Array also contains the Hamasaki-Cube main layer signals, which have been shown to exhibit a slight increase of HF-content compared to the Hybrid-Array main layer signals. In any case, the results regarding this attribute stand in contradiction with the claim that one of the main advantages of the Main-Array would be its ability to convey Presence [13].

#### 4.2.7. Timbral Colouration

Dummy head recordings have been conducted to gather an objective reference for comparison to the timbral qualities perceived in the auditory evaluation. However, no direct indications could be derived from the spectral graphs derived thereof about the nature of the possible timbral colourations of the different arrays. It is proposed that an experimental design with fewer uncontrolled variables should be applied to approach this complex matter, similar as in [12].

## 5. Conclusions

This work has been useful in gaining an understanding of the spatial and timbral perception of a Bowles-Array, a Fukada-Tree/Hamasaki-Cube configuration and their hybrid version. The report of the different experimental techniques and the discussion of their outcomes gave further indications on how individual array parts might have contributed to the perception of specific attributes. This insight could be seen as a basis for recording engineers experimenting with 3D recording techniques for informing some of their decisions.

## 6. References

- [1] Bowles, D.: A microphone array for recording music in surround-sound with height channels, presented at the 139th International AES Convention (2015 October), preprint 9430.
- [2] Gribben, C. and Lee, H.: The Frequency and Loudspeaker-Azimuth Dependencies of Vertical Interchannel Decorrelation on the Vertical Spread of an Auditory Image. *Journal of the Audio Engineering Society* (2018), 66(7/8), 537-555.
- [3] Hamasaki, K. and Van Baelen, W.: Natural Sound Recording of an Orchestra with Three-dimensional

- Sound, presented at the 138th International AES Convention (2015 May), preprint 9348.
- [4] Hanyu, T., Kimura, S. and Chiba, S.: A New Objective Measure For Evaluation of Listener Envelopment Focusing on the Spatial Balance of Reflections. *Journal of Architecture and Planning* (1999), 64(520), 9-16.
- [5] Holman, T.: *Surround Sound: Up and Running*, 2nd ed., Focal Press, Oxon, 2008.
- [6] Howie, W. and King, R.: Exploratory microphone techniques for three-dimensional classical music recording, presented at the 138th International AES Convention (2015 May), e-Brief 196.
- [7] Howie, W., King, R. and Martin, D.: A Three-Dimensional Orchestral Music Recording Technique, Optimized for 22.2 Multichannel Sound, presented at the 141st International AES Convention (2016 September), paper 9612.
- [8] Howie, W., Martin, D., Benson, D., Kelly, J. and King, R.: Subjective and objective evaluation of 9ch three-dimensional acoustic music recording techniques, presented at the AES International Conference on Spatial Reproduction - Aesthetics and Science (2018 July), convention paper P10-1.
- [9] Kassier, R., Lee, H., Brookes, T. and Rumsey, F.: An Informal Comparison Between Surround-Sound Microphone Techniques, presented at the 118th International AES Convention (2005 May), preprint 6429.
- [10] Lee, H., Gribben, C. and Wallis, R.: Psychoacoustic Considerations in Surround Sound with Height, presented at the 28th Tonmeistertagung VDT International Convention (2014 November).
- [11] Riaz, H., Stiles, M., Armstrong, C., Chadwick, A., Lee, H. and Kearney, G.: Multichannel Microphone Array Recording for Popular Music Production in Virtual Reality, presented at the 143rd International AES Convention (2017 October), e-Brief 384.
- [12] Robotham, T., Stephenson, M. and Lee, H.: The Effect of a Vertical Reflection on the Relationship between Preference and Perceived Change in Timbral and Spatial Attributes, presented at the 140th International AES Convention (2016 May), preprint 9547.
- [13] Theile, G. and Wittek, H.: 3D Audio Natural Recording, presented at the 27th Tonmeistertagung VDT International Convention (2012 November).
- [14] Wallis, R. and Lee, H.: The Reduction of Vertical Interchannel Crosstalk: The Analysis of Localisation Thresholds for Natural Sound Sources. *Applied Sciences* (2017), 7(3), 278-297.
- [15] Zielinsky, G., Lemmens, P., Dabringhaus, W., Wittek, H. and Nettingsmeier, J., Proceedings of ICSA 2011. DVD. VDT, Detmold, 2011.
- [16] Berg, J. and Rumsey, F.: Identification of Quality Attributes of Spatial Audio by Repertory Grid Technique. *Journal of the Audio Engineering Society* (2006), (5), 365-379.
- [17] Berg, J. and Rumsey, F.: Systematic Evaluation of Perceived Spatial Quality, presented at the 24th AES International Conference (2003 June), paper 43.
- [18] Gerzon, M.: *Whither Four Channels?.* Link House Publications, Croydon, 1971.
- [19] Rumsey, F.: Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm. *Journal of the Audio Engineering Society* (2002), 50(9), 651-666.
- [20] Pedersen, T. and Zacharov, N.: The development of a Sound Wheel for Reproduced Sound, presented at the International AES Convention (2015 May), preprint 9310.
- [21] Gribben, C. and Lee, H.: Towards the Development of a Universal Listening Test Interface Generator in Max, presented at the 138th International AES Convention (2015 May), e-Brief 187.
- [22] Howie, W., King, R., Martin, D. and Grond, F.: Subjective Evaluation of Orchestral Music Recording Techniques for Three-Dimensional Audio, presented at the 142nd International AES Convention (2017 May), preprint 9797.
- [23] Rumsey, F.: *Spatial Audio*. Focal Press, Oxon, 2013.
- [24] Bech, S. and Zacharov, N.: *Perceptual Audio Evaluation - Theory, Method and Application*. Wiley, Chichester, 2006.
- [25] Rumsey, F., Zieliński, S., Kassier, R. and Bech, S.: On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality. *The Journal of the Acoustical Society of America* (2005), 118(2), 968-976.
- [26] Toole, F.: Subjective Measurements of Loudspeaker Sound Quality and Listener Performance. *Journal of the Audio Engineering Society* (1985), 33(1/2), 2-32.
- [27] Berg, J. and Rumsey, F.: Correlation between Emotive, Descriptive and Naturalness Attributes in Subjective Data Relating to Spatial Sound Reproduction, presented at the 109th International AES Convention (2000 September), preprint 5206.
- [28] Martin, B., King, R. and Woszczyk, W: Subjective Graphical Representation of Microphone Arrays for Vertical Imaging and Three Dimensional Capture of Acoustic Instruments, Part I, presented at the 141st International AES Convention (2016 September), preprint 9613.
- [29] Barron, M.: The subjective effects of first reflections in concert halls - The need for lateral reflections. *Journal of Sound and Vibration* (1971), 15(4), 475-494.





# Ambisonic Decoder Description (.ADD)

Presented \* by VDT.

## Developing a new file format for ambisonic decoding matrices

G. Arlauskas, J. Ohland, H. Schaar

*University of Applied Sciences Darmstadt, Germany,*

*Email: {jonas.ohland, gabriel.p.arlauskas, henning.schaar}@stud.h-da.de*

### Abstract

Different software solutions have been developed for the calculation and implementation of ambisonic decoding matrices. The present paper presents and describes a new data file format which can be used as an intermediate between solutions.

Currently available software solutions use particular data conventions causing difficult compatibility and exchangeability. In the present work an open-source toolkit is developed for storing, handling and using ambisonic decoding matrices. The toolkit includes tools for conversion from common matrix data conventions to the ADD-format and back, calculating decoding matrices, decoding ambisonic signals and extracting existing matrices from external decoding tools.

The new ADD-format and toolkit enables increased flexibility in production workflows and eliminates the drawbacks and limitations regarding compatibility between software solutions.

## 1. Introduction

Spatial audio can be considered an extension of surround sound but in addition to the horizontal plane, the whole three dimensional sound field is described. Ambisonics has been established as a reliable mathematical way to represent the sound field components [1]. Those components are obtained by encoding a sound source with spherical harmonics.

Spherical harmonics are an infinite set of harmonic functions defined over the surface of a sphere and can be defined as

$$Y_\ell^m(\vartheta) = N_{\ell,|m|}^X P_\ell^{|m|}(\cos \theta) \begin{cases} \cos(m\phi), & \text{for } m \geq 0 \\ \sin(|m|\phi), & \text{for } m < 0 \end{cases} \quad (1)$$

where  $\vartheta$  is the angular direction,  $P$  is the associated Legendre polynomial of order  $\ell$  and index  $m$ , and  $N$  is a normalisation factor obtained by a method X defined either for the Schmidt semi-normalized (SN3D) as

$$N_{\ell,|m|}^{\text{SN3D}} = \sqrt{\frac{2 - \delta_m (\ell - |m|)!}{4\pi (\ell + |m|)!}} \delta_m \begin{cases} 1 & \text{if } m = 0 \\ 0 & \text{if } m \neq 0 \end{cases} \quad (2)$$

or for the fully normalized form (N3D) with an additional factor

$$N_{\ell,|m|}^{\text{N3D}} = N_{\ell,|m|}^{\text{SN3D}} \sqrt{2\ell + 1} \quad (3)$$

And the encoding of a set of  $k$  input signals  $g_l$  can be expressed as

$$\hat{\phi} = \sum_{k=1}^K \mathbf{y}(\vartheta_k) g_k \quad (4)$$

where

$$\mathbf{y}(\vartheta) := [Y_0^0(\vartheta), \dots, Y_\ell^m(\vartheta), \dots]^T$$

or in a simplified form as

$$\hat{\phi} = \Upsilon \mathbf{g} \quad (5)$$

where

$$\begin{aligned} \mathbf{g} &:= [g_1, \dots, g_K] \\ \Upsilon &:= [y(\vartheta_1), \dots, y(\vartheta_K)] \end{aligned}$$

In most cases the signals  $\mathbf{s}$  decoded for various loudspeaker setups can be obtained by applying a decoding matrix  $D$

$$\mathbf{s} = D\phi \quad (6)$$

Thus decoding an ambisonic signal, as long as the position of the speakers is constant, is a static operation with low complexity once the matrix has been calculated.

In recent years, numerous approaches have been presented to calculate these matrices. Above all, the approaches differ in their suitability regarding certain speaker systems and playback situations and contents. A good overview of existing approaches, their advantages and disadvantages can be found in [2] and [3].

Often these methods are difficult to use in practice. Many of the methods described exist only as implementations for applications tailored to specific fields of research.

For example, decoding matrices well suited for irregular loudspeaker layouts can be calculated with implementations created for Matlab, such as the EPAD [4], CSAD [5] and AllRAD [6] method. However, their use in applications like *ambidecode~* [7] in Max/MSP proves to be difficult. Although both solutions support importing and exporting of decoding matrices as files, the formats are not compatible with each other, even though they both contain the same data.

To overcome compatibility problems like these, we propose the *.add* format. It should serve as a bridge between different solutions, while still providing enough flexibility to incorporate all common features as well as future developments.

## 2. Method

In order to design such a format, firstly it is recommendable to get an overview of the requirements that meet existing formats. Obviously, it is necessary to describe at least one decoding matrix  $D$  which transfers an incoming set of SH  $\phi$  to the reproduction channels  $\mathbf{s}$ . Especially interesting though, is the additional data produced by the applications we analyzed, such as e.g. speaker positions, normalization method and output routing.

In this context, we investigated the following solutions:

**ambidecode~** *ambidecode~* developed in the Zurich University of Arts [8] and described in [7] allows to export and

import the internal matrix as an xml file. In addition, the expected normalization of the incoming ambisonic signal as well as the layout of the speaker system, a gain factor for each output, and a set of decoding weights for the ambisonic components can be exported to a separate file.

**Ambix decoder** In addition to the matrix, the configuration files for the ambix decoder [9] also contain information about the expected order of spherical harmonics and a gain factor for the entire matrix.

**Compact higher-order Ambisonic Library** This is a collection of Matlab functions for use with higher order Ambisonics [10]. None of these functions were explicitly written to generate files but this can be done easily.

**IEM AllRAD decoder** The AllRAD decoder [11] exports not only the matrix and some metadata such as a name and a description but also information about the expected normalization of the ambisonic signal, a desired weighting of ambisonic components per order, and the layout of the target rendering system.

**IEM Simple Decoder** The Simple Decoder [11] is unable to produce a matrix itself but can read files produced by AllRAD Decoder and use the matrices it contains for ambisonic decoding. In addition, one can specify a subwoofer channel, which will be taken from the matrix output and passed through a high pass filter after decoding.

**Ambilibrium** Ambilibrium [12] exports configuration files for the Ambix decoder and the format used by the IEM.

**AmbDec** AmbDec [13] enables the separation of an ambisonic signal into two frequency bands and to then decode it with different decoding matrices. Thus, the format produced by the AmbDec offers the possibility to store two matrices, a crossover frequency and a relative gain factor for both frequency bands. Information about the expected normalization, the speaker layout and whether the decoder should make a latency compensation when all speakers are not on the surface of a sphere can also be represented.

**Ambisonics decoder toolbox** The Ambisonics decoder toolbox [14] exports *ambdec* files and configuration files for the ambix decoder.

## 3. Design

For designing the *.add* format we decided to stay close to the design of the configuration files for the IEM plugins. Additionally, an optional filter stage, optional additional matrices and extended metadata can be saved. All filters and all outputs can be named. An *.add* file contains a creation date, author information, details about the software it was created with, and a version number. To avoid compatibility problems, a format revision is saved in each file.

Papers written by Heller et al. [15, 16] describe the advantages of multiband decoding. For this reason, the .add format supports the basic description of filters applied to the ambisonic signal before the decoding step with corresponding matrices.

In the description of the filters, we have decided to restrict ourselves to specifying the cut-off frequencies and to leave the filter design to the implementation. However, we encourage the use of phase-matched IIR Filters to preserve uniform frequency response over all directions.

Channel ordering in .add files is done according to the ACN standard, where the channel number can be determined algorithmically:

$$ACN = \ell^2 + \ell + m \tag{7}$$

The expected type of normalization is specified with each file and may either be SN3D or N3D as in (2) and (3) respectively.

As a container format, we chose JSON for a variety of reasons. JSON is widely used and supported by many programming languages. The structure of a JSON object can be represented by native constructs in those languages which can be manipulated intuitively. Furthermore, it is human-readable and can be edited with a simple text editor. Compared to XML, JSON strings are favorable regarding data storage space.

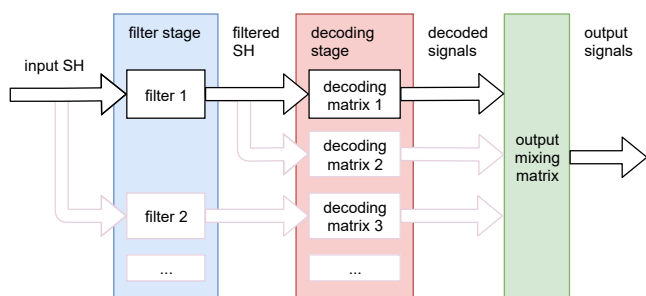


Fig. 1: Schematic representation of an ambisonic decoder that can be described by the .add format.

## 4. Implementation

**dotaddtool** The dotaddtool converts the various formats for storing decoder matrices into .add files and back. This tool enables end users to work with multiple softwares, which have not implemented the .add file format themselves. As such it also serves as an intermediate solution, before .add files have seen widespread adoption.

**Decoder Extractor** Another part of the toolkit is the "dotadd-decoder-extractor" that aims for the particular case when a plugin doesn't allow exporting of decoder matrices.

The tool consists of two VST Plugins using a peer-to-peer inter-process connection which are placed up- and downstream right next to a target decoder plugin. Since a matrix is applied passively in the processing algorithm, sending a signal with a value of 1 through the processor for every channel

results in the output being solely the internal decoding matrix used in the plugin.

Because matrix data can vary depending on e.g. Ambisonics order or channel configuration, it is possible to configure the output signal of the exciter plugin to produce fitting and functional matrix data. The values of the matrix are cached on run-time, recalculated according to configurations specified by the user and lastly exported as an .add-file. The tool can only extract matrices from decoders that do not apply any additional filtering or band-splitting.

**Software libraries** In order to guarantee the uniformity of .add files and to facilitate adaptation, software libraries for the programming languages C ++, Python, JavaScript and Matlab are developed. These allow fast adaptation of existing program infrastructure to the .add format.

Fig. 2 shows a complete example of creating an .add file describing a basic ambisonic decoder with a single output.

```

1  const ADD = require("dotadd.js");
2  const fs = require("fs");
3
4  let add_file = new ADD()
5      .setName("Example Decoder")
6      .setAuthor("My Name");
7
8  add_file.addMatrix(new ADD.Matrix(
9      [[1., 0., 0., 0.]));
10
11 fs.writeFileSync("/output/file.add",
12     add_file.export().serialize());

```

Fig. 2: Creation and export of an .add file in JavaScript for the node.js runtime environment.

The Software libraries and tools will be released soon and can be found under <https://github.com/smp-3d/dotadd>.

## 5. Discussion

Along the history of Ambisonics, developers have contributed to the technology in a gradual manner making the theory practically accessible. Provision of full-fledged toolkits such as SPARTA [17], the IEM PluginSuite [18] or the ICST Ambisonics Externals for Max/MSP [8] has pushed the limits for widespread usage of Ambisonics technology. However, compatibility issues are still common and prevent industrial use of multi-software solutions. As for decoders it still lacks a common ground to improve further according to agreed upon standards. Our proposal is a further step towards a universal workflow with this type of technology. We are optimistic about the outcome and are curious about improvements and open for conversation.

## 6. Acknowledgements

We would like to thank Dr. Jorge Medina for his expert advice and ongoing encouragement. We would also like to thank Prof. Thorsten Greiner and Prof. Felix Krückels for providing the environment without which this work would not have been possible.

## 7. References

- [1] M. A. Gerzon. 1985. Ambisonics in multichannel broadcasting and video. *J. Audio Eng. Soc.*, 33, 11, 859–871. <http://www.aes.org/e-lib/browse.cfm?elib=4419>.
- [2] F. Zotter, M. Frank, and H. Pomberger. 2013. Comparison of energy-preserving and all-round ambisonic decoders. *Journal of the Audio Engineering Society*, 60, (January 2013), 807–820.
- [3] F. Zotter, H. Pomberger, and M. Noisternig. 2010. Ambisonic decoding with and without mode-matching: a case study using the hemisphere. In *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics* (Paris, France). (April 2010).
- [4] F. Zotter, H. Pomberger, and M. Noisternig. 2012. Energy-preserving ambisonic decoding. *Acta Acustica United with Acustica*, 98(1), 37–47.
- [5] N. Epain, C. T. Jin, and F. Zotter. 2014. Ambisonic decoding with constant angular spread. In .
- [6] F. Zotter and M. Frank. 2012. All-round ambisonic panning and decoding. *Journal of the Audio Engineering Society*, 60, (October 2012), 807–820.
- [7] J. C. Schacher and P. Kocher. 2006. Ambisonics spatialization tools for max/msp. In *Proc. of the 2006 International Computer Music Conference*, 274–277.
- [8] Zurich University of Arts. 2019. Ambisonics externals for max/msp. Retrieved 08/24/2019 from <https://www.zhdk.ch/forschung/icst/software-downloads-5379/downloads-ambisonics-externals-for-maxmsp-5381>.
- [9] M. Kronlachner. 2014. ambiX v0.2.8 - ambisonic plug-in suite. Retrieved 08/24/2019 from <http://www.matthiaskronlachner.com/?p=2015>.
- [10] A. Politis. 2015. Compact higher-order ambisonic library. Retrieved 08/30/2019 from <http://research.spa.aalto.fi/projects/ambi-lib/ambi.html>.
- [11] IEM - Institute of Electronic Music and Acoustics. 2019. IEM Plugins configuration files manual. Retrieved 08/30/2019 from <https://plugins.iem.at/docs/configurationfiles/>.
- [12] M. Romanov. 2018. Ambilibrium - a user-friendly ambisonics encoder/decoder-matrix designer tool. In *Audio Engineering Society Convention 144*. (May 2018). <http://www.aes.org/e-lib/browse.cfm?elib=19541>.
- [13] F. Adriaensen. 2019. Ambdec user manual - 0.4.2. Retrieved 08/30/2019 from <https://kokkinizita.linuxaudio.org/linuxaudio/downloads/ambdec-manual.pdf>.
- [14] A. Heller. 2014. The ambisonic decoder toolbox: extensions for partial-coverage loudspeaker arrays. In *Proc. of Linux Audio Conference 2014*.
- [15] A. Heller, R. Lee, and E. Benjamin. 2008. Is my decoder ambisonic? In *Audio Engineering Society Convention 125*. (October 2008). <http://www.aes.org/e-lib/browse.cfm?elib=14705>.
- [16] J. Heller A. and E. M. Benjamin. 2018. Design and implementation of filters for ambisonic decoders. In *Proc. of the 1st International Faust Conference (IFC-18)* (Mainz, Germany). (July 2018).
- [17] L. McCormack. 2018. Spatial audio real-time applications (SPARTA). Retrieved 08/24/2019 from [http://research.spa.aalto.fi/projects/sparta\\_vsts/](http://research.spa.aalto.fi/projects/sparta_vsts/).
- [18] IEM - Institute of Electronic Music and Acoustics. 2019. IEM Plug-in Suite. Retrieved 08/30/2019 from <https://plugins.iem.at>.



# Full Reviewed Paper at ICSA 2019

Presented \* by VDT.

## A Systematic Performance

### Investigation of Convolution Algorithms for Synthetic Room Reverberation

S. Heidtmann<sup>1</sup>, W. Fohl<sup>2</sup>

<sup>1</sup> HAW Hamburg, Germany, Email: stefan.heidtmann@haw-hamburg.de

<sup>2</sup> HAW Hamburg, Germany, Email: wolfgang.fohl@haw-hamburg.de

#### Abstract

This paper presents a systematic investigation of optimization strategies for the convolution algorithm. Special attention is given to features relevant for the creation of virtual room acoustics, where the source signal is convolved with a room impulse response signal which has a length of several seconds. Examined were optimizations for the discrete convolution in the time domain and for the partitioned fast convolution in the frequency domain. Applied technologies were usage of AVX instructions, and GPU computing with the OpenCL framework. The results of the various algorithms are evaluated in terms of sample throughput. Various influence factors on the measured performance were identified. It turned out, that even ambitious projects with more than 10 channels and filter response lengths of several seconds may be rendered in real-time with the GPU version of the discrete convolution.

## 1. Introduction

Convolution is one of the key algorithms in digital audio processing. FIR filters perform the convolution of the input signal with the filter's impulse response. A special application is the creation of virtual room acoustics in multichannel setups: the original signal is convolved with several directional room impulse responses; the resulting audio signal is played back from the proper directions so that listeners get a realistic impression of the room reverberations.

If the acoustic environment created by this method is supposed to be *interactive*, the virtual room reverberations must be calculated in *real-time*. In order to achieve this goal, several challenges have to be met. Room impulse responses are considerably longer than the FIR filter lengths usually employed in audio processing. Typical reverb durations are in the range of several seconds, resulting in filter lengths of some 100,000 samples.

Approximately 24 reverberation signals will be needed for each primary source to create a realistic room reverb in a WFS system [1].

For interactive environments, *latency* is another important feature. Furthermore, the interaction may cause changes to the virtual

room acoustics. These changes must also be recognizable with low latency.

The implementation of the discrete convolution in time domain has an algorithmic complexity of  $\mathcal{O}(n^2)$ , so for large virtual rooms an optimized implementation of convolution algorithms is of crucial importance. Therefore a systematic investigation of optimization strategies for convolution algorithms has been conducted, taking into account modern techniques as AVX and GPU computing as well as classical code optimization.

The rest of this paper is organized as follows: In the following section related work is given, then the covered optimization strategies are presented. After that the test environment is described and the results are presented and discussed.

## 2. Related Work

Artificial reverberations is a relatively old topic. The first real-time reverberations were realized in hardware and later by digital filter structures. The real-time usage of convolution reverb is a development over the last 10 years that was enabled by the increasing

power of CPUs and the usage of GPUs for the convolution [2].

That a GPU can be used to increase the performance of different audio processing tasks was shown by L.Savioja, V.Välämäki, J.O. Smith [3]. They showed that with the CUDA framework, a GPU can be used to increase the performance of additive Synthesis, discrete and fast convolution. A problem in performance testing is that the result depends on the hardware. Since the hardware, and the compiler for said hardware, improves with time, the results of older papers may not reflect the current state. Similar results were reported by Wefers and Berg [4].

One of the more recent studies focused on the performance difference between executing the fast convolution on the CPU, showing a performance advantage for the GPU for larger problem sizes [5].

Though most papers focus on the fast convolution, the discrete convolution is of particular interest for real-time audio applications, not only because of their better performance for certain problem sizes. An issue with real-time audio processing is the latency between input and output. The latency depends on the size of the processing buffer, a smaller buffer means a shorter latency. When an effect is added to an instrument in real-time, this latency becomes noticeable. How noticeable depends on the instrument. The acceptable range can range from 1.4 to 42 milliseconds [6]. The processing buffer size for 1.4ms latency would be 62 samples for a sample rate of 44.1 kHz or 68 samples for 48 kHz.

The properties of the discrete convolution, except the algorithmic complexity, are better for audio processing than the properties of the fast convolution. The size of the processing buffer does not affect the processing time of the discrete convolution, so low latencies are easily achievable, and replacing the filter response takes up to no time. The better performance for smaller problem size makes it potentially a better solution for some problems. Other approaches to the latency problem are hybrid convolution [7] and non-uniformly-partitioned convolution [8]. The Hybrid convolution algorithm only convolves the direct sound and the early reflection and use other methods to create the remaining reverberations. Non-uniformly partitioned convolution convolves the signals for the smaller partitions and calculates the larger ones with the fast convolution.

### 3. Optimization Strategies

Convolution algorithms have a high algorithmic complexity, but the choice of the algorithm is not the only influencing factor on the performance. A major contributing factor is how well optimized the code is and on what kind of device the code is executed. In this document three approaches to increase the performance of the convolution algorithm by reducing the execution time are evaluated. The first approach is the optimization of the code by standard optimization techniques. The second is the usage of intrinsic functions, functions that call processor specific operations, to improve the performance by using the SIMD unit of a CPU. The last approach is the usage of the GPU of a PC through OpenCL.

**Code Optimization** Code optimization can be achieved by various means. The options include: reducing the number of CPU operations, replacing slow operation through faster ones, reducing the memory transfer between RAM and CPU Caches as well as between Caches and the CPU Registers.

**Operator Replacing** Replacing of slow operation can drastically reduce the execution time of code. Two of the slowest arithmetic operation are the *division* and *modulo* operation, but both can be replaced under certain circumstances.

If many divisions by the same divisor occur, it is more efficient to compute once the inverse of the divisor, and subsequently replace the divisions by the multiplication with the inverse.

Replacing the modulo operation is only possible in integer arithmetic, when both operands are positive and the right operand is a *power of two*. If these conditions are met, the modulo operation can be replaced by a bitwise AND:

$$x \% 2^n = x \& (2^n - 1) | x, n \in \mathbb{N}$$

**Loop unrolling** Loop unrolling is common practice to reduce the number of operations and jumps in the code, since jumps are costly operations. Loop unrolling means to reduce the number of iteration by executing the loop body multiple times in a single iteration. This does not only reduce the number of jumps, but also the number of compare operations [9].

**Register Optimization** When more variables are used in a code block than the CPU has registers, *register spilling* can occur: The CPU has to store the content of the registers into the cache to free up space for further operations. This slows down the execution time. Usage of the registers can only be directly controlled in assembler but by reducing the number of currently used variables the compiler can optimize the register usage [9].

**Advanced Vector Extensions** AVX improves the performance of uniform operations on array elements by processing multiple data at the same time. AVX instructions are either used directly in assembler code or by using intrinsic functions in C/C++. AVX can also be used by the compiler if enabled but there is no guarantee that the compiler will use it [9].

**OpenCL** The Open Computing Language is a framework for developing and executing programs on different platforms for parallel computing, mainly on the GPU, but also on the CPU, on FPGA and DSP [10]. OpenCL differentiates between two components, the *host device* and the *computing device*. The host device acts as a master that starts the execution of OpenCL programs, the so-called *kernels*, and controls the memory transfer between the devices. Each kernel contains a task for the computation of *one single* result element. The kernels are written in Open CL C, a programming language that is based on the syntax of C.

OpenCL code can be optimized by using the same techniques as described previously. To further optimize the performance, the programmer has to properly make use of global and local memory. A common optimization strategy for the memory is *tiling*. In tiling a problem is divided into multiple smaller parts, where each part is small enough that it can be executed in a single working group of cores [11, 12, 13].

### 4. Convolution Engine Implementations

The discrete and the fast convolution algorithm were implemented multiple times. Each of them multiple times in C++ and OpenCL usually in an unoptimized variant and an optimized variant to show the benefits of the optimizations. While the implementations of the convolution engine interface differ in the used hardware,

```

1  int n = currentRingBufferPos;
2  for (int i = n; i < I/O_Buffer.size; i++, n++){
3      for (int m = 0; m < Filter.size; m++){
4          I/O_Buffer[i] +=
5              Save_Buffer[(n - m) % Save_Buffer.size] * Filter[m];
6      }
7  }

```

**Fig. 1:** Pseudocode for the discrete convolution

and the used algorithm, the internal buffer structure is similar. All of them have a buffer for the filter response, the I/O buffer, and a buffer to hold intermediate data.

The convolution algorithms are explained with the help of pseudocode. To improve readability, the pseudocode describes only the convolution of a single channel. The iteration over the different audio channels as well as the normalization of the audio are missing. The normalization is simply a multiplication of all samples in the result and the iteration over the audio channels an additional loop and index for all buffer accesses.

### 4.1. Discrete Convolution Engine

This engine implements the discrete convolution using the Overlap Save approach. This implementation was created to have an unoptimized implementation as a reference of the discrete convolution for comparison with other engines.

The discrete convolution engine implements the convolution by implementing the equation 1 to calculate a sample for the result.

$$(f * g)[n] = \sum_{m=0}^{|g|} f[n-m] \cdot g[m] \quad (1)$$

The convolution engine implementation needs two `for` loops for processing an I/O buffer (Fig. 1). The first loop iterates over the samples in the I/O buffer and the second loop implements the sum in the equation.

The implementation of the discrete convolution is the implementation of the equation 1 with only a small modification in form of a modulo operation (Fig. 1, line 4). The modulo operation is necessary because the `Save_Buffer` is a ring buffer. The access to the `Save_Buffer` in line 4 runs backwards. Through the modulo operation the access jumps from the lowest element of the buffer to the highest.

For correct functionality, the ring buffer has to be able to hold at least the same amount of data than the sum of frame size and length of the filter response. This size is necessary because every sample of the filter response is multiplied with a sample in the `Save Buffer` from a starting point in the reverse direction (Fig. 1). A ring buffer is used in every convolution engine for Overlap Add as well as for Overlap Save.

#### 4.1.1. Optimized Discrete Convolution

Optimizations for reducing the execution time of code are usually applied at the expense of the readability and maintainability of the code (Fig. 2). The most efficient way to optimize code is to optimize the parts of the code that are executed the most, meaning mostly the bodies of loops.

The convolution operation of a single channel in the convolution

```

1  size_t moduloMask = Save_Buffer.size - 1;
2  int n = currentRingBufferPos;
3  for (int i = 0; i < Frame.size; i += 8) {
4      for (int j = 0; j < Filter.length; j += 4) {
5          float filter_0 = Filter[j];
6          // ... etc.
7          float filter_3 = Filter[j + 3];
8
9          float val0 = Save_Buffer[(n + i - j) & moduloMask];
10         // ... etc.
11         float
12             val3 = Save_Buffer[(n + i - j - 3) & moduloMask];
13
14         I/O_Buffer[i] += val0 * filter_0 + val1 * filter_1
15
16             + val2 * filter_2 + val3 * filter_3;
17         // ... etc
18         val1 = Save_Buffer[(n + i - j - 7) & moduloMask];
19         I/O_Buffer[i + 7] += val1 * filter_0 + val2 * filter_1
20
21             + val3 * filter_2 + val0 * filter_3;
22     }
23 }

```

**Fig. 2:** Pseudocode for the optimized version of the discrete convolution. Shorter and more readable than the actual implementation

```

1  for (int i = 0; i < Frame.size; i += 8){
2      size_t moduloMask = Save_Buffer.size - 1;
3      vecResult = loadValue(0);
4
5      for (int j = 0; j < Filter.size; j++){
6          vecIn = load(&Save_Buffer[(i - j) & moduloMask]);
7          vecFilter = loadValue(Filter[j]);
8          vecResult += vecIn * vecFilter;
9      }
10     store(&I/OBuffer[i], vecResult)
11 }

```

**Fig. 3:** Pseudocode for the discrete convolution using vector instructions

engine was optimized by *loop unrolling* of the outer loop (Fig. 2, line 3) and the inner loop (Fig. 2, line 4). To avoid register spilling, the number of local variables was reduced by cyclic changing of the `val` variables (Fig. 2, line 9, 11, 16, etc.).

The last optimization visible in the pseudocode, was to replace the modulo operation with a bitwise AND (Fig. 2, line 9, 11, 16). For this optimization to work, the size of the ring buffer is rounded up to the next power of two.

#### 4.1.2. AVX Discrete Convolution

The vector arithmetic unit of a CPU allows operations on a 256-bit vector. This allows to add or multiply eight floats at the same time. The vectors can be initialized by loading data from a float array (Fig. 3, line 3) and can also be written into a float array (Fig. 3, line 10). The AVX implementation always calculates eight samples for the result at the time. The result is calculated by loading a block of eight samples from the `Save_Buffer` and multiply them with a single value of the filter response. The result of the multiplication is then added to a result register (Fig. 3, line 5 - 7).

The main advantage of AVX is the potentially higher throughput through the use of vector instructions. Additionally, the header of the second loop is executed less often since eight samples are processed at the same time, the same effect as loop unrolling (Fig. 3, line 2).

For clarity, the wrap around of the ring buffer is not shown in the pseudocode.



```

1 copy(transform_time , last_I/O_Buffer)
2 copy(&transform_time[transform_time/2] , I/O_Buffer)
3
4 fft(transform_time , transform_freq)
5 copy(fd1[currentPartition] , transform_freq)
6
7 int n = currentPartition + nrOfPartition
8 for(int j = 0; j < nrOfPartition; j++)
9     for (int i = 0; i < transform_freq.size; i++){
10         freq_accumulation
11         [i] = filter_freq[j][i] * fd1[n - j][i];
12     }
13
14 ifft(freq_accumulation , transform_time)
15 copy( I/O_Buffer , &transform_time[transform_time/2]);
16
17 currentPartition = (currentPartition + 1) % nrOfPartition
18 last_I/O_Buffer = I/O_Buffer
    
```

Fig. 4: Pseudocode for the fast convolution with uniform partition

### 4.2. Uniformly Partitioned Engine

The algorithm with uniformly partitioned filter responses is a variant of the fast convolution that is specifically suited for the block-wise processing in audio applications. This algorithm outperforms in all respects the unpartitioned fast convolution algorithm, where the transformed audio signal is multiplied with the filter response in one single step. So the unpartitioned fast convolution is not discussed further.

The algorithm of the uniformly partitioned engine starts with copying the last I/O buffer into the first half of a transform buffer and the current I/O buffer into the second half (Fig. 4, line 1-2). The transform buffer is then transformed into the frequency domain and its content is put into a FDL (Frequency-domain delay line) (Fig. 4, line 4-5). The convolution is then carried out by multiplying the filter partitions with the entries in the FDL and adding them in an accumulation buffer (Fig. 4, line 7-12). The last step is to transform the accumulation buffer into the frequency domain and copy the second half into the IO buffer (Fig. 4, line 14-15).

A particular optimization is the access to the FDL ring buffer. The required modulo operation could be replaced by an AND operation if the size of the FDL would be increased to a power of two, but instead the size of the FDL is doubled. The first half of the FDL are pointers to the buffers storing the frequency data. The second half is equal to the first half. Since the implementation iterates backward over the entries the implementation starts in the second half. When the wrap around would happen the implementations simply enters the first half.

A version using AVX for the multiplication of the complex numbers exists as well.

### 4.3. Multithreading

All modern CPU have multiple independent cores. Using them is an effective approach to increase the sample throughput, but additional time is needed for the synchronization of the threads. The multithreading in the convolution engines was designed to minimize the synchronization overhead by assigning the convolution of an audio channel to a single thread. The channels are equally distributed to the threads. Because each thread fully utilizes the computing power of a core, using more threads than there are CPU cores does not increase the sample throughput.

The advantage of this approach is, that the threads are completely

```

1 convolve(Frame, Save_Buffer, Filter, n){
2 int channel_id = get_global_id(0);
3 int sample_id = get_global_id(1);
4
5 channel_save = SaveBuffer[channel_id]
6 channel_filter = Filter[channel_id]
7
8 float result;
9 for (int m = 0; m < Frame.Size; m++){
10     result +=
11         channel_save[(n + sample_id - m) % channel_save.size]
12         * channel_filter[m];
13 Frame[channel_id][sample_id] = result;
14 }
    
```

Fig. 5: Pseudocode for the computation device for the discrete convolution with OpenCL

independent from the other threads during the convolution. Synchronization is only needed to start the threads and to wait until all threads have completed their task.

### 4.4. OpenCL Implementations

Convolution engines using the discrete and the partitioned convolution have been implemented for Open CL. The OpenCL implementations try to reduce the involvement of the CPU to a minimum. The only task of the CPU is the memory transfer and calling of the kernels. The convolution itself is only carried out on the GPU.

In OpenCL, the optimizations can be categorized into two categories: Optimization of the *kernel code* and optimization of the *memory transfer* between the devices. OpenCL is supported by a range of devices, but the optimizations of OpenCL code were applied to maximize the performance on a GPU. The applied optimization of the code may lead to worse performance on other device types.

#### 4.4.1. Discrete Convolution

The simple implementation of the discrete convolution with the GPU is similar to the implementation of the discrete convolution on the CPU (Fig. 5). Like the CPU version the GPU version implements Overlap Save. The major difference is that the loops of the samples in the I/O buffer is missing. The loop is implemented by spawning a GPU thread for every iteration of the loop. The amount of threads spawned is controlled by the host device (Fig. 6, line 8).

The kernel of all threads is the implementation of the equation for the discrete convolution (eq. 1). The equation calculates a single value for the result. The thread gets the information which channel and sample they have to calculate by the id of the thread. In OpenCL a thread has a three-dimensional id. In this case, the first dimension is the index of the audio channel and the second the sample in the result that the thread has to calculate (Fig. 5, line 2-3). The last dimension is not used.

#### 4.4.2. Kernel Optimization

The main difference between the standard discrete convolution kernel and the optimized version is the usage of *tiling*. Tiling is a technique to improve the sample throughput by dividing a calculation into smaller tiles to make use of the local memory of the device.

In OpenCL, each thread is part of a work group. A work group on the GPU consists of multiple GPU cores for parallel code execution and a shared local memory. The local memory is smaller but faster than the global memory of the GPU. It is comparable to the caches in the CPU and is shared by all threads.



```

1 cl_cmdQueue.writeBuffer(Frame, FrameCL)
2
3 cl_kernel.setArg(0, CL_Frame);
4 cl_kernel.setArg(1, CL_Save_Buffer);
5 cl_kernel.setArg(2, CL_Filter);
6 cl_kernel.setArg(3, currentRingBufferPosition);
7
8 cl::NDRange global(Number of Channel, Frame.Size);
9 cl_cmdQueue.callKernel(convolve, global);
10
11 cl_cmdQueue.readBuffer(Frame, FrameCL)

```

**Fig. 6:** Pseudocode for the host for executing the discrete convolution with OpenCL

The number of cores in a working group varies from one device to another. On GPU it is usually 32 or 64 cores [11, 12, 13].

Because the local memory is limited, larger problems like the convolution have to be divided into multiple tiles. In practice, the processing of the frame buffer is divided into multiple tiles with either 32 or 64 samples each. This is generally like splitting the frame buffer into multiple smaller buffers. Each working group fully processes one tile (Fig. 7) and like the unoptimized version, every thread calculates one sample for the result [11, 12, 13].

For fast transfer between local memory and global memory the loading process of the thread in a working group needs to be aligned (Fig. 7, line 8 - 13). Aligned means that when a thread with the id  $x$  transfers element  $x$  from local to global memory the thread with id  $(x + 1)$  has to do same for the element  $(x + 1)$ . If the transfer is aligned, the transfer is a single instruction, if not, the GPU needs one instruction for every single transferred value [11, 12, 13].

Read access to the Save Buffer and the filter response is realized in blocks. The threads load a tile sized block of data into the local buffer and then access the local buffer to calculate the result (Fig. 7, line 12, 18, 19). At all times two blocks of the save buffer have to be loaded to correctly calculate the result. The kernel uses a ring buffer to allow this.

#### 4.4.3. Memory Transfer

The convolution can be optimized by a better usage of the involved hardware, mainly the PCI-E bus. The execution time required to process a frame can be divided into three parts (Fig. 9): Data transfer to the OpenCL device, execution of the convolution, and data transfer to the host. This means that during the time the device processes the current frame the memory bus is idling and vice versa.

By changing the host code the available hardware can be better used by implementing a pipeline (Fig. 8). Processing the frame buffer needs three frames. During the first frame the input frame is transferred to the device, in the second frame the buffer is processed, and in the last frame, the processed data is transferred from the device to the host. This means that at any given time three frame buffers are in the pipeline.

The advantages of the pipeline are, that the time for processing a frame depends only on the longest execution time for one of its components and thus increasing the sample throughput. This is only the case when the memory bus is either full duplex or dual simplex, like PCI-E.

There are two disadvantages, namely a latency between in- and output of three full frames and more memory is needed on the comput-

```

1 convolve(Frame, Save_Buffer, Filter, n,
2   local save, local filter, tile_size){
3   float result = 0;
4   int channel_id = get_global_id(0);
5   int tile_id = get_global_id(1);
6   int sample_id = get_global_id(2);
7
8   saveIndex = n + sample_id + tile_id * tile_size;
9   filterIndex = sample_id;
10  bufferPointer = 0;
11
12  save[bufferPointer] = SaveBuffer[channel_id][saveIndex];
13  bufferPointer = (bufferPointer + tile_size) % save.size;
14
15  for(int i; i = 0; i < Filter.Size / tile.Size; i++){
16    saveIndex =
17      (saveIndex - tileSize) % SaveBuffer[channel_id].Size;
18
19    save
20      [bufferPointer] = SaveBuffer[channel_id][saveIndex];
21    filter
22      [bufferPointer] = Filter[channel_id][filterIndex];
23    bufferPointer
24      = (bufferPointer + tile_size) % save.size;
25
26    for (int m = 0; m < tile_Size; m++){
27      result
28        += save[(bufferPointer + sample_id - m) % tile_size]
29          * filter[m];
30    }
31    filterIndex += tileSize;
32  }
33  Frame[channel_id]
34    [[sample_id + tile_id * tile_size] = result;
35  }

```

**Fig. 7:** Pseudocode for an optimized version of the discrete convolution with OpenCL. For this code to work all buffer have to have a size that is a multiple of the tile\_size

```

1 temp = outputBuffer;
2 outputBuffer = processingBuffer;
3 processingBuffer = inputBuffer;
4 inputBuffer = outputBuffer;
5
6 cl_cmdQueue.writeBuffer(FrameCL[inputBuffer], Frame)
7 cl_cmdQueue
8   .readBuffer(FrameTempOut, FrameCL[outputBuffer])
9
10 cl_kernel.setArg(0, FrameCL[processingBuffer]);
11 cl_kernel.setArg(1, CL_Save_Buffer);
12 cl_kernel.setArg(2, CL_Filter);
13 cl_kernel.setArg(3, currentRingBufferPosition);
14
15 cl::NDRange global(Number of Channel, Frame.Size);
16 cl_cmdQueue.callKernel(convolve, global);

```

**Fig. 8:** Pseudocode for the host for executing the discrete convolution with a pipeline approach for memory transfer and device computation

ing device. While the latency can be compensated when the size of the I/O buffer could be reduced to a third of the size than otherwise possible, the increased memory consumption can not be circumvented. The cause for the increase in memory consumption is, that the content of a frame buffer is filled by the host while the device needs them for executing kernel. Because of this three frame buffer are needed on the GPU that are switched by the host (Fig. 8, 1-4).

#### 4.4.4. Uniformly Partitioned Convolution

The OpenCL partitioned convolution implementation uses the CLFFT library for better performance of the Fourier transforms on the GPU. The host code differs from the host code of the discrete convolution. Instead of using a single kernel, three are used to convolve the signal (Fig. 10). One kernel is for the transform, one for the multiplication of the frequencies and one for the inverse

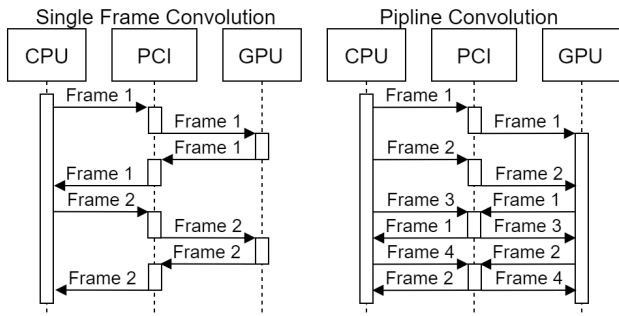


Fig. 9: Pipeline Models for the Open CL Implementation

```

1 copy(transform_time , last_I/O_Buffer)
2 copy(&transform_time[transform_time/2] , I/O_Buffer)
3
4 cl_cmdQueue.writeBuffer(transformBuffer , transform_time)
5
6 cl_kernel.setArg(fftargs);
7 cl::NDRange global(Number of Channel , Frame.Size);
8 cl_cmdQueue.callKernel(transform , global);
9
10 cl_cmdQueue.copy(fd1[currentPartition] , transformBuffer)
11
12 cl_kernel.setArg(complex multiplication);
13 cl::NDRange global(Number of Channel , Frame.Size);
14 cl_cmdQueue.callKernel(cmplx_mult , global);
15
16 cl_kernel.setArg(iffargs);
17 cl::NDRange global(Number of Channel , Frame.Size);
18 cl_cmdQueue.callKernel(inverse_transform , global);
19
20 cl_cmdQueue.readBuffer(
21     I/O_Buffer , &transformBuffer[transformBuffer.size/2])
22
23 currentPartition = (currentPartition + 1) % nrOfPartition
24 last_I/O_Buffer = I/O_Buffer
25

```

Fig. 10: Pseudocode for the host for executing the partitioned convolution

transform.

In the implementation the kernel carries out the convolution by multiplying the filter partitions with the entries, while the FFT and the IFFT are provided by CLFFT. The task of the host is to move the data around, like moving the data from the transform buffer into the FDL (Fig. 10, line 10).

## 5. Results and Discussion

In this chapter measurement results for the various engines are reported. All measurements were carried out on a Intel Core i7-4770 CPU with 4 cores and a maximum clock frequency of 3.9 GHz and 16 GB RAM. The GPU is a NVIDIA GTX 970 with a maximum clock frequency of 1.316 GHz, 4 GB memory and 13 computing units.

The standard parameter set for the convolution measurements is: *I/O buffer* of 256 samples, *filter length* of 48,000 samples, and 4 *simultaneous channels*.

Only one of these parameters has been varied in the measurements.

### 5.1. Performance Comparison Between Discrete and Fast Convolution

The plot in Fig. 11 shows the sample throughput of a discrete convolution engine and a fast convolution engine for different processing buffer sizes for all power of two buffer sizes between

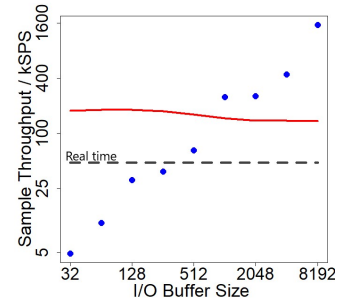


Fig. 11: Sample throughput for the discrete convolution (red) and the fast convolution (blue) for different processing buffer sizes with a filter length of 440,100 Samples. The 48,000 samples per second line is marked by a dashed grey line.

32 and 8192. As seen the in the plot the discrete convolution is mostly unaffected by the change in the buffer size.

On the other hand the sample throughput of the fast convolution grows exponentially with the size of the processing buffer. The discrete convolution has a throughput of roughly 200,000 samples, while the fast convolution starts with 5000 at a buffer size of 32 break the 44,100 mark at a buffer size of 512 and overtake the discrete convolution at a buffer size of 1024. This also means, that unlike the discrete convolution, the convolution of two signals need more time than convolving a single signal with twice the filter length.

The fast convolution data is shown as dots instead of a line in Fig. 11, because the performance of the FFT varies heavily with the length of the transform array. Lengths that are large prime numbers give very poor performance, many small prime factors are good, ideal are powers of two.

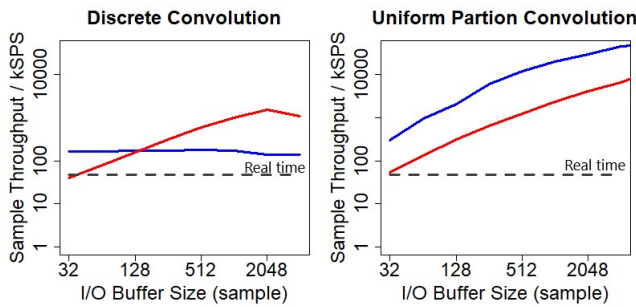
### 5.2. Performance Comparison Between CPU and GPU Implementations

This section presents the results for the most performant implementations of the discrete and partitioned fast convolution, on the CPU and the GPU for various values of I/O buffer size, filter length, and number of channels.

#### 5.2.1. I/O Buffer Size

The first engine parameter tested is the I/O buffer size. While theoretically the size of the I/O buffer does not matter for the performance of the discrete convolution, the GPU performance changes when the I/O buffer size changes (Fig. 12). The lowest sample throughput for the discrete convolution on the GPU is with 40 thousand SPS at an I/O buffer size of 32 samples too low for real-time processing. With increasing buffer size the sample through of this convolution engine increases through to better efficiency of the memory transfer. The sample throughput peaks at a buffer size of 2048 with a throughput of 1524 kSPS. On the CPU the sample throughput is in comparison relatively constant. The peak is at an I/O buffer size of 512 with 175 kSPS and at its lowest at a buffer size of 4096 with a throughput of 136 kSPS. A comparable performance between the two convolution engines is at a buffer size of 128 where the CPU version has with 172 kSPS a slightly higher throughput than the OpenCL version, with 155 kSPS.

For the partitioned convolution the behavior is slightly different (Fig. 12). Like the discrete convolution the OpenCL



**Fig. 12:** Sample throughput of the convolution engines by varying size of the I/O buffers. Start value for the buffer size were 32 samples, end 4096 measurement points were all power of twos in between. In this and the subsequent plots, the CPU implementations are in blue, the OpenCL implementations in red, and the 48 kHz threshold for processing audio in real-time for the common sample rates is indicated by the dashed gray line.

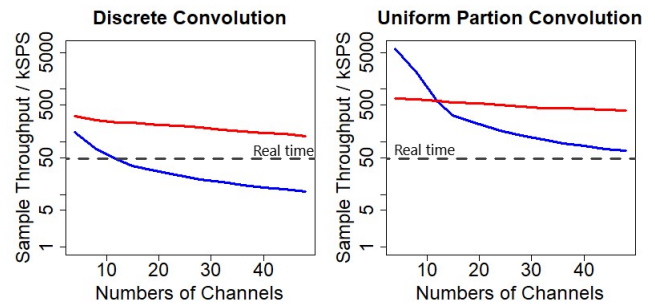
implementation of the partitioned convolution starts lower than the CPU implementation with a sample throughput of 52 kSPS, but unlike the discrete convolution the OpenCL implementation is not able to surpass the CPU implementation. The CPU implementation starts with a sample throughput of 295 kSPS, a value that is surpassed by the OpenCL implementation at a buffer size of 128 samples. The throughput of both implementations increases with increases in the I/O buffer size. At the last measured buffer size of 4096 the OpenCL implementation has a throughput of 6.6 mSPS, and the CPU version an throughput of 45 mSPS.

**5.2.2. Number of Channels**

The next parameter to be examined is the number of parallel convolutions. The plots show the behavior of the convolution engines when the number of channels is increased (Fig. 13). The tests start at four channels, are incremented in four channels steps and finally end at a channel number of 48. All convolution engines start with a higher throughput than necessary for a sample rate of 48kHz. No engine with the exception of the CPU discrete convolution fall below this threshold, but the CPU partitioned convolution at channel number 48 is only slightly above the real-time limit with a throughput of 51.5 kSPS. The CPU discrete convolution falls under the 48 kHz threshold when convolving 12 channels. The sample throughput at this point is 46,8 kSPS, less than a third of the sample throughput for four channels 153.5kHz. This sample throughput is too low for a sample rate of 48 kHz but just sufficient for convolving 12 channels at a sample rate of 44.1 kHz.

The sample throughput of the convolution engines decreases with an increase in the number of channels for the OpenCL implementations in a somewhat linear fashion (red line in Fig. 13). On the other hand, the CPU partitioned convolution (blue line) first drastically loses performance and then slows down. The sample throughput of this engine decreases fast from 5.92 mSPS to 585 kSPS at a channel size of 12, the break-even point between the CPU and the OpenCL version.

It is noteworthy, that the performance of the OpenCL partitioned convolution varies only slowly with channel number: The performance at 60 channels is only smaller than the performance at 4 channels by a factor of 1.9.



**Fig. 13:** Sample throughput of the convolution engines by a varying number of channels for the convolution. The start number of channels is 4, and the end 48. Test were executed between start and end in increments of four. Blue: CPU, red: GPU

**5.2.3. Filter Length**

The last parameter to be examined is the length of the filter. In a first experiment the performance of the discrete and partitioned convolution are measured for rather long filter lengths of 24,000 to 480,000 samples, corresponding to filter durations from 0.5 s to 10 s at 48 kHz sampling rate. The results are shown in Fig. 14.

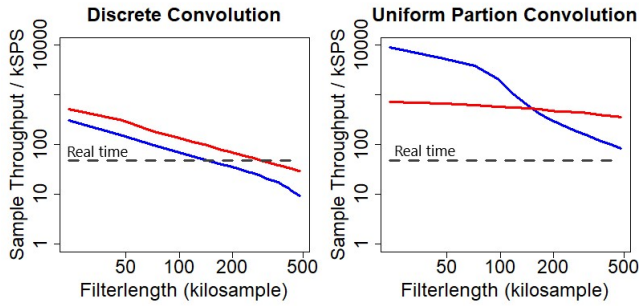
In a second test, the measurements are repeated for small problem sizes. The results are given in Fig. 15. The goal of these experiments is to find out if there is a break-even point between discrete and partitioned fast convolution. As the plot shows, the uniformly partitioned convolution faster than the discrete convolution even for the smallest filter lengths.

The plots in figure 14 shows that the engine behaves similarly with regard to the filter length as to the number of channels. The OpenCL implementation of the discrete convolution is yet again always better than the CPU implementation, but this time both end up under the 48 kHz threshold. The CPU engine after a filter length of 144,000 samples the OpenCL engine at 288,000 samples, coincidentally the OpenCL discrete convolution engine reaches twice as many samples.

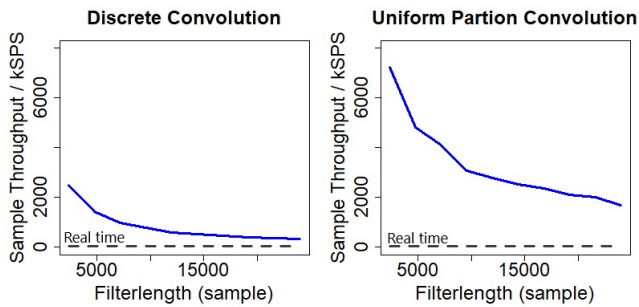
The CPU partitioned convolution outperforms the OpenCL version drastically for shorter filter lengths, but breaks even with it between 144,000 and 168,000 samples. The point when the CPU version falls below the real-time rate of 48,000 samples is reached outside of the scope of the plot at a filter length around 720,000.

The largest filter length that can be processed with our current OpenCL implementation is 5,760,000 samples. At this point, the sample throughput is still 50,000 SPS. This filter length is equal to 120 seconds of audio at a sample rate of 48 kHz.

That the partitioned convolution is not only good for long filter length shows another plot showing the sample throughput for smaller filter lengths for the CPU implementations (Fig. 15). As always the case the partitioned convolution outperforms the discrete convolution. The value range of the discrete convolution is between 25,000 kSPS and 322 kSPS, while the value of the partitioned convolution ranges from 7250 kSPS to 1686 kSPS. This clearly shows that the uniform partition also handles relatively small filter lengths far better than the discrete convolution.



**Fig. 14:** Sample throughput of the convolution engines by varying filter length. The start filter length is 24,000, and the end 480,000. Tests were executed between start and end in increments of 24,000. Blue: CPU, red: GPU



**Fig. 15:** Sample throughput of the CPU convolution engines for short filter lengths. The start filter length is 2400, and the end 20,600. Tests were executed between start and end in increments of 2400. The size of the I/O buffers is in this case 64.

## 6. Conclusion

Optimization experiments have been conducted to measure the performance of various convolution implementations. The base parameters were chosen for a typical scenario of virtual room acoustics rendering at 48 kHz: I/O buffer of 256 samples, 4 simultaneous channels, filterlength of 48 kSamples (corresponding to 1 second of room reverb). From this base setting one of the three parameters was varied and the influence on performance was observed. Compared were optimized CPU and GPU implementations in time and frequency domain (discrete convolution and uniformly partitioned fast convolution).

In all experiments, the performance of the frequency-domain implementation was considerably higher, even for small I/O buffers (32 samples), or short filterlengths (2400).

The result of the CPU / GPU comparison is, that the overhead of the GPU implementations in time and frequency domain only pay off at larger parameter sizes. There is one remarkable exception: The CPU-based frequency-domain convolution with 4 channels and 480 kSamples filterlengths is considerably faster at all tested I/O buffer sizes (32 to 4096).

There are however situations in interactive applications, where it is necessary to modify the filter response during runtime. These cases are more easily realized using the discrete convolution. Our results show, that a properly optimized GPU-based *discrete* convolution algorithm is able to handle filterlengths of about 100 kSamples and an I/O buffer size of 256 for some dozens of channels in realtime.

As a proof of concept a VST plugin was developed which allows the selection of the various convolution algorithms and filter responses. With this plugin it was possible to render the acoustics of the WDR concert hall [1] with 24 channels at 48 kHz sampling rate in real-time for our 208-channel WFS system.

## 7. References

- [1] Wolfgang Fohl and Eva Wilk. Enhancements to a wave field synthesis system to create an interactive immersive audio environment. In Proc. 3rd Int. Conf. on Spatial Audio. VDT, 2015
- [2] V. Välimäki, J. D. Parker, L. Savioja, J. O. Smith, and J. S. Abel. Fifty years of artificial reverberation. IEEE Transactions on Audio, Speech, and Language Processing **20** (2012), 1421–1448. ISSN 1558-7916
- [3] Lauri Savioja, Vesa Välimäki, and Julius O. Smith. Audio signal processing using graphics processing units. J. Audio Eng. Soc **59** (2011), 3–19
- [4] F. Wefers and J. Berg. High-performance real-time fir-filtering using fast convolution on graphics hardware. In 13th International Conference on Digital Audio Effects (DAFx-10). Graz, Austria, 2010
- [5] D. V. Nikolov, M. J. Mišić, and M. V. Tomašević. Gpu-based implementation of reverb effect. In 2015 23rd Telecommunications Forum Telfor (TELFOR). 2015 990–993
- [6] Michael Lester and Jon Boley. The effects of latency on live sound monitoring. In Audio Engineering Society Convention 123. 2007
- [7] A. Primavera, S. Cecchi, F. Piazza, J. Li, and Y. Yan. Hybrid reverberator using multiple impulse responses for audio rendering improvement. In 2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing. 2013 314–317
- [8] N. Jillings, J. D. Reiss, and R. Stables. Zero-delay large signal convolution using multiple processor architectures. In 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). ISSN 1947-1629, 2017 339–343
- [9] Intel Corporation. Intel 64 and IA-32 Architectures Optimization Reference Manual, 2018
- [10] Aaftab Munshi, Benedict Gaster, Timothy G Mattson, and Dan Ginsburg. OpenCL programming guide. Pearson Education, 2011
- [11] John L Hennessy and David A Patterson. Computer architecture: a quantitative approach. Elsevier, 2011
- [12] Nvidia Corporation. OpenCL Programming Guide for the CUDA Architecture, 2009
- [13] Advanced Micro Devices Corporation. OpenCL User Guide, 2015





## Full Reviewed Paper at ICSA 2019

Presented \* by VDT.

### Auralization systems for simulation of augmented reality experiences in virtual environments

Peter Dodds, Sebastià V. Amengual Garí, W. Owen Brimijoin, Philip W. Robinson  
*Facebook Reality Labs, Redmond, WA, USA, Email: peterdodds@fb.com*

#### Abstract

Augmented reality has the potential to connect people anywhere, anytime, and provide them with interactive virtual objects that enhance their lives. To deliver contextually appropriate audio for these experiences, a much greater understanding of how users will interact with augmented content and each other is needed. This contribution presents a system for evaluating human behavior and augmented reality device performance in calibrated synthesized environments. The system consists of a spherical loudspeaker array capable of spatial audio reproduction in a noise isolated and acoustically dampened room. The space is equipped with motion capture systems that track listener position, orientation, and eye gaze direction in temporal synchrony with audio playback and capture to allow for interactive control over the acoustic environment. In addition to spatial audio content from the loudspeaker array, supplementary virtual objects can be presented to listeners using motion-tracked unoccluding headphones. The system facilitates a wide array of studies relating to augmented reality research including communication ecology, spatial hearing, room acoustics, and device performance. System applications and configuration, calibration, processing, and validation routines are presented.

#### 1. Introduction

We imagine a future world in which a multitude of people will wear augmented reality devices that overlay an entire metaverse of auditory and visual information on their daily experiences. Such devices will influence the way we live, work, and interact with other people. These devices will give us super-human listening abilities and facilitate telepresence with a much higher social signal bandwidth. To reach this future, we will need a much greater understanding of how users with augmented abilities will interact with the devices and each other.

Consider an example where two people are sitting in a noisy restaurant talking to one another. Devices with motion tracking, simultaneous localization and mapping, a microphone array, and binaural sound synthesis could provide many benefits in this situation. With the relative positions and orientations of the participants known, one listener's microphone array

could capture the other talker with reduced background noise, and play it back in real time, binaurally positioned in the same spatial location as the actual talker, hence naturally and transparently improving the signal to noise ratio of transmitted speech. The shape and dynamic tuning of the beamformer, the allowable delay of the reinforcement signal, the accuracy of the spatialization, and the values of many other parameters needed to optimize the performance of the system in this scenario are all unknown.

The technology imagined in the scenario above largely exists today, just not in a form factor that is suitable for a head-worn, mobile device. Many technological developments are still needed to package such capabilities into a consumer product. Nonetheless, without having these future devices available for testing, we must utilize what is available to prototype the experiences and further research in the area. This paper describes a real-time interactive auralization system that utilizes

currently available technology to prototype future experiences to better understand how users with augmented abilities will interact with their devices and each other.

## 2. System Overview

Our interactive auralization platform (IAP) is an evolving combination of hardware and software, designed to allow us to test new hardware and software, examine novel AR experiences, and rigorously measure human behavior in real and augmented reality environments. The facility allows us to test, evaluate, and demonstrate the various technologies that are being developed by the FRL audio research group, as well as assess human behavior in multiple subjects simultaneously in a variety of realistic acoustic environments, both real and with virtual components to them. In this way, it acts as a time machine because some of the functionality is not yet possible in current generations of AR hardware, even those that exist only in prototype form. The IAP has a high density loudspeaker array capable of reproducing realistic sound fields and is also capable of presenting real-time spatialized virtual audio wirelessly over headphones, including virtual sound sources, shared audio from nearby talkers, and remote talkers (telepresence), all rendered with very low latency using individualized head-related transfer functions (HRTFs), simulated room acoustics, and a virtual beamformer. The way the IAP is constructed gives us an ability to rapidly iterate changes in functionality to determine the utility and ideal parameters for a potential device feature without having to do extensive hardware development.

The IAP is also capable of making fine-grained measurements of the behavior of multiple people interacting in a common or physically separated space. These measurement capabilities currently include head, hand, and body tracking, eye gaze and pupillometry, and high-fidelity voice capture for up to six simultaneous subjects. Future additions will include finger and face tracking, galvanic skin responses measures, electroencephalography, and heart rate. All the measurement subsystems are tied to a common master clock (Evertz 5601MSC), allowing accurate time synchronization of current and future measurement data. This is critical for examining realtime functionality of devices with novel sensors and capabilities and also analyzing complex behaviors that are associated with real and virtual events in the space. The ability to capture data from multiple people at the same time ensures that we can capture and utilize details of interactive communication, exploration, and play in groups of people.

## 3. Subsystem Details

### 3.1. Loudspeaker Array

#### 3.1.1. Hardware

In order to best understand how users will interact with future augmented reality devices, the wide variety of environments in which the devices will be used must be examined. To this end, a semi-spherical loudspeaker array for spatial audio reproduction has been constructed in the interactive auralization platform. The array consists of 49 MiniDSP SPK-4P

loudspeakers. The SPK4-P is a compact loudspeaker with a 3.5" driver and an on-board 400 MHz Analog Devices SHARC processor and Class D amplifier. A single CAT5/6 network cable from a PoE or PoE+ enabled switch provides power, audio, and control using the Audio Video Bridging (AVB) communication protocol. For full-band reproduction, audio signals below 120Hz are sent to four miniDSP NDAC-2 AVB endpoints which convert the AVB stream to analog audio to drive four Genelec 7360A SAM Studio Subwoofers.

The loudspeaker array is interfaced to a single Windows PC computer in an isolated control room using an RME Digiface USB. Max/MSP and Matlab are used for real-time processing of the audio streams which allows the array to use different audio rendering pipelines simultaneously and interface with the other render and capture technologies in the system. Figure 1 illustrates the signal path of the IAP subsystems, including the loudspeaker array and audio pipeline.

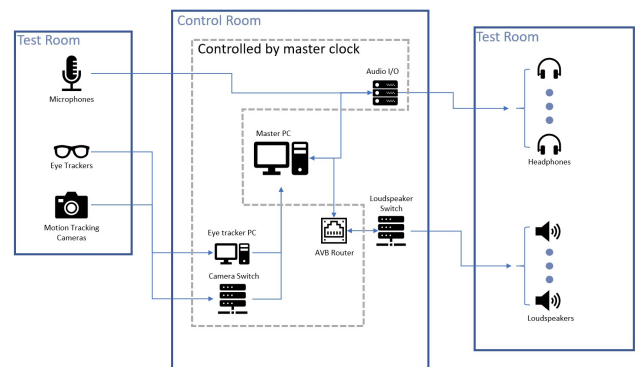


Fig. 1: Schematic diagram of IAP hardware connections.

#### 3.1.2. Array Implementation

Figure 2 shows the 53 loudspeaker array for sound field reproduction that has been installed as part of the Interactive Auralization Platform. The positions of the loudspeaker were determined by circumscribing a sphere on to the room and then optimizing the positions of the loudspeaker for spatial audio reproduction given the architectural constraints of the room. The array consists of four rings of 12 loudspeakers each, roughly approximating a sphere, and a single loudspeaker directly above the center of the room. The subwoofers are placed at the cardinal compass directions at the edge of the room.

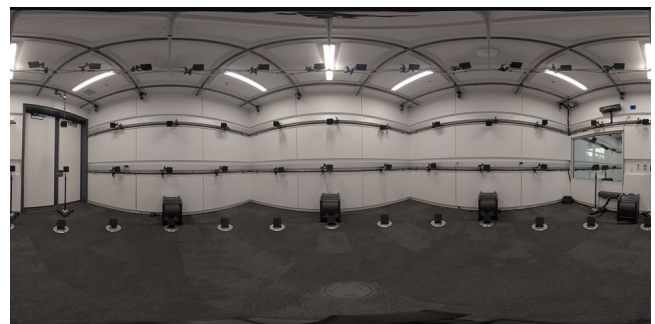
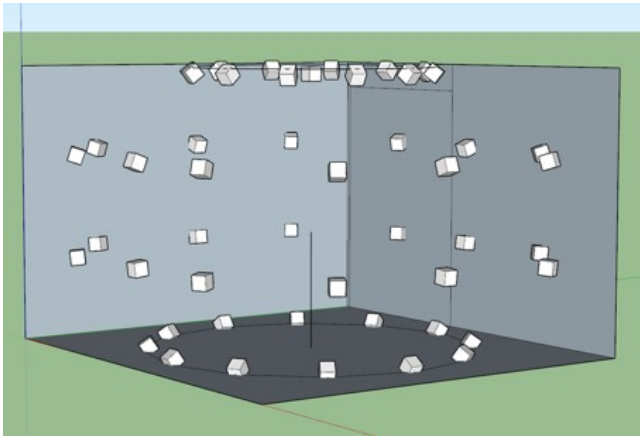


Fig. 2: Equirectangular photo of one of the loudspeaker arrays built for the IAP. (Photo credit: Scott Colburn)

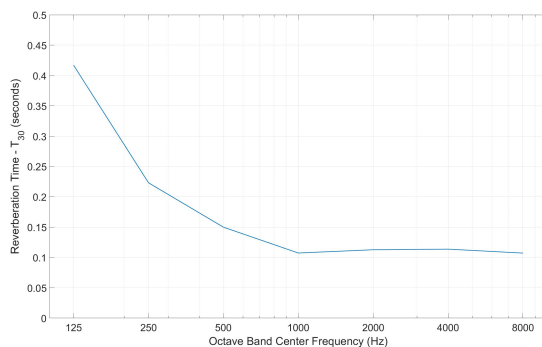
The loudspeakers are individually calibrated with respect to level and delay and corrected to minimize coloration of the soundfield using an automated Matlab script. These filters can be loaded on to the SHARC processor of each loudspeaker or implemented further up the signal path, in the computer, depending on the needs of the IAP. Using the software back-end, a variety of spatial reproduction formats are achievable over the loudspeakers, including vector based amplitude panning (VBAP) [1], Spatial Decomposition Method (SDM) [2], and Higher Order Ambisonics (up to order  $N = 6$ ).



**Fig. 3:** 3D model showing the loudspeaker positions within the room.

**3.1.3. Room Acoustics**

In order to reduce the effects of room reverberance on the soundfield reconstruction, the room is acoustically treated. Two inch absorptive paneling is placed on all major surfaces of the room, save for an observation window and the floor. The frequency-dependent reverberation time of the room is shown in Figure 4.



**Fig. 4:** Measured reverberation time  $T_{30}$  in the IAP.

**3.2. Motion Tracking**

The IAP uses a Vicon motion tracking system, which employs retroreflective infrared markers that may be attached to people or objects and can be tracked at sub-millimeter resolution throughout the capture space. The system currently in use consists of eight Vicon Bonita 10 cameras, mounted at ceiling height in the cardinal compass directions, capable of running at a sample rate of 250 Hz. An alternative version of the

IAP uses Optitrack Prime 13W cameras running at 120 Hz. The motion tracking systems may be used to capture data on listener movements for behavioral capture but is also used to drive a real-time loudspeaker array and binaural rendering system. The position and rotation data of all objects tracked in the space is accessed using the Vicon DataStream SDK (or Optitrack Motive SDK), which streams the Cartesian coordinates and 4-element quaternions of each tracked object to other pieces of software on the same or other computers on the subnet. Currently this data is captured in Matlab, which uses a custom script to compute the angles (azimuth and elevation) and distance of all tracked objects relative to each other, in local coordinate frames. Untracked virtual objects may also be added at this stage. The relative Euler angles and distances are then rebroadcast at a frame rate of 100 Hz over UDP to Max/MSP, which handles the room acoustic rendering, loudspeaker DSP, and/or binaural spatialization.

**3.3. Eye Tracking**

Gaze tracking is accomplished using Ergoneers glasses (Dikablis II) containing infrared emitters and three cameras: two eye-facing cameras and one world-facing camera, all running at a frame rate of 60 Hz. These devices can be wireless or wired, depending on configuration, and interface with our data acquisition and streaming pipeline with D-Lab, with data being shared over UDP to Vicon. This allows us to assess and utilize gaze angle to alter sound presentation parameters in realtime in the same coordinate reference frame as the motion tracking data, and further allows us to ensure that data timestamps are uniform across the different software packages and measurement systems. Data captured for offline analysis includes gaze direction, pupil height, and pupil width, with HD videos captured from all three cameras for further analysis.

**3.4. Voice Capture**

Audio is captured for analysis or re-spatialization from each participant using DPA microphones (4088 directional microphone) in a boom-mount configuration. The signals are transmitted over wireless transmitters (Sennheiser SK 100 G4) and captured in Max/MSP using an RME Fireface UFX II multi-channel sound card at a sample rate of 48 kHz. Time alignment between presented and captured signals is ensured by using the same sound card, which along with the motion and eye tracking systems is slaved to the IAP master clock.

**3.5. Headphone Subsystem**

Virtual sound sources may be presented binaurally over headphones in the IAP, but care must be taken to minimize the impact that the headphones have on the listener's perception of real world or loudspeaker array-generated signals. The ability to present sounds to a listener without interfering with the natural sound path is a fundamental requirement of auditory AR. As there are no currently available headphones that allow the presentation of fully broadband binaural signals to completely unoccluded ears without cross-talk, three types of commercially available

devices are used in our work, varying inversely in the width of their frequency response and in the degree to which they impede the signal path of real world signals into the ear canal. The first are AKG K1000 open ear headphones, large diaphragm headphones that are suspended over the ears, rather than being pressed against the head. These are high fidelity headphones and have the strongest low frequency response of the headphones used, but they interfere the most with the natural sound field due to their large size. The second type that is used are Sony PFR-v1 headphones, small, spherical loudspeakers suspended in front of the pinnae, with a bass port that is routed through a hollow metal loop that sits in the ear behind the tragus. These are capable of a relatively flat response over a wide range of frequencies but have poorer low frequency response than the AKGs as they are smaller, however, they interfere less with the natural sound field. Finally, a pair of custom headphones are used that consist of a small in-ear headphones (Sennheiser IE4) that are suspended roughly 1 cm from the opening of the ear canal using a custom 3D printed mount. These have the poorest low frequency reproduction, but in principle occlude the ears the least. Two other categories of AR transducers are also currently used, but as they are proprietary technology they will not be discussed here. All headphone signals are presented via an RME Fireface UFXII sound card either wired or wirelessly with remote monitoring transceiver (Sennheiser EW IEM G4).

## 4. Audio Stimuli

### 4.1. Sound Generation

All audio is processed in Max/MSP, utilizing a variety of pre-made and custom objects including the SPAT toolbox from IRCAM. Max/MSP receives the azimuth, elevation, distance, and level of the desired virtual sound sources via UDP from Matlab and renders the appropriate virtual signals at these locations for up to six listeners. Arbitrary HRTFs can be loaded into the rendering software, ensuring flexibility in presenting individualized signals to multiple listeners. Soundfields are rendered using pre-recorded ambisonics content or by artificially generating sources using traditional sound design techniques and encoding them in to a spatial reproduction format. Room acoustics simulations are handled in two different ways, depending on the configuration of the system, but both are implemented in Max/MSP.

### 4.2. Room Acoustics Simulation

#### 4.2.1. Room Re-Synthesis

Generating perceptually plausible virtual acoustic objects in AR requires matching the acoustics of the simulated percepts to those of the real room. Once the acoustic divergence between real and virtual sounds becomes too large, virtual binaural percepts do not appear well externalized, and thus impair the quality of experience [3]. However, it is unknown what are acceptable deviations and to what extent it is necessary to match the acoustics of the virtual sounds to those of the real space. To enable research in perceptual thresholds of room acoustics for augmented reality it is useful to generate

binaural renderings that are based on the measured acoustics of the room.

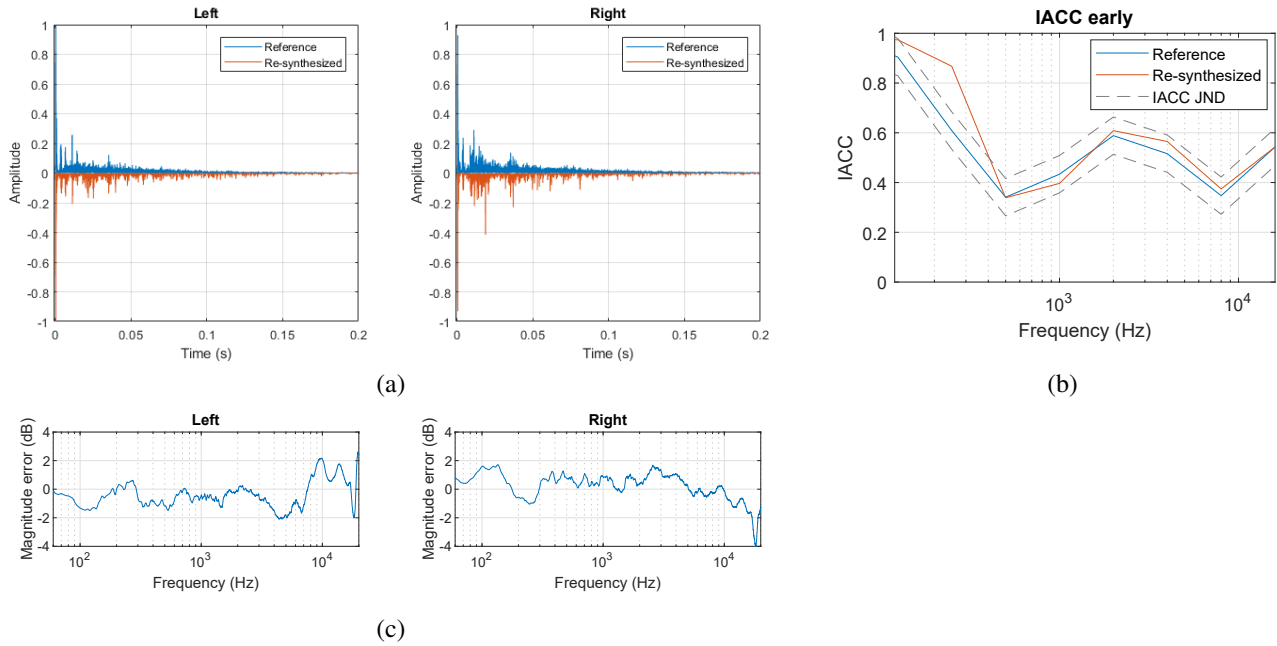
A straightforward implementation consists of measuring a binaural room impulse response (BRIR) using a head and torso simulator (HATS) with a variety of head orientations. However, this approach is time consuming and does not allow for personalization using individualized HRTFs. An alternative approach is based on microphone array measurements at the listening position and binaurally re-synthesizing the sound field. The method used here is largely based on the Spatial Decomposition Method (SDM) and the re-synthesis of BRIRs by combining the sound-field parametrization data (a pressure monaural RIR and direction-of-arrival information) with an arbitrary dataset of HRTFs. An extensive technical description and validation of the auralization approach is described in [4]. A variety of manipulations of the rendered BRIRs is presented as well in [4], including arbitrary modifications of the frequency-dependent reverberation time, the inclusion of fully synthetic late reverberation, or the manipulation of the spatial characteristics of the sound field.

The re-synthesized BRIRs are dynamically convolved in real-time in three separate pipelines: direct sound, early reflections, and late reverberation. This allows the implementation of real-time manipulations of the BRIR, enabling the study of HRTF manipulations, direct-to-reverberant ratio (DRR), or mixing time, among others. Reproducing room acoustics based on in situ measurements over non-occluding headphones allows us to compare real and virtual sources in real-time and objectively assess the perceptual differences between them.

Pilot listening tests have been completed to assess the degree of authenticity and plausibility of the auralizations. A discrimination test based on a two-alternative forced choice (2AFC) with a reference revealed that if listeners are provided with unlimited listening time the binaural renderings are not indistinguishable from the real loudspeaker. To our knowledge, there is only one study available in the literature testing perceptual authenticity of dynamic binaural synthesis [5]. In that case, although the BRIRs were measured in situ using binaural microphones, perceptual authenticity was not achieved. We have found that typically small deviations in spectral content and localization are the most common attributes used to judge deviations between a reference real sound and a binaural rendering. In order to test plausibility, we conducted a pilot study where a loudspeaker was covered behind an acoustically transparent baffle, and spatially degraded versions of a binaural render were compared to the real loudspeaker. In this case, we found that real and virtual sources appear to be equally plausible. Formal studies are being conducted to confirm these initial findings.

Figure 5 shows a comparison between a measured BRIR with a mannequin and a re-synthesized version of this BRIR using the above described method. As it can be observed, the time-energy properties of the left and right channels are largely preserved, although some spurious reflections





**Fig. 5:** (a) Absolute value of a BRIR measured with a mannequin and a re-synthesis. (b) Interaural Cross Correlation of the early part (0 to 80 ms) of the measured and re-synthesized BRIRs. (c) Spectral error after monaural equalization.

can be observed in the re-synthesis. The Interaural Cross Correlation (IACC) of the measured and re-synthesized versions fall within  $\pm 1$  JND (0.075 as defined in the standard ISO 3382). The spectral error falls within  $\pm 2$  dB up to 16 kHz.

#### 4.2.2. Artificial Reverberation

In this configuration, flexibility and realtime performance is prioritized, and is used when participants are expected to walk around the room, interacting with multiple virtual and real sound sources. Here we use the feedback delay network in SPAT, with the parameters manually matched to the natural acoustics of the IAP room. Alternative room models may also be used that do not match the acoustics of natural signals in the room. In the future, the real-time propagation engine found in the Oculus Audio SDK will be implemented as an additional room simulation configuration.

### 5. Example Use Cases of the IAP

#### 5.1. Virtual spatial mixing of musical compositions

In this scenario, two participants are fitted with wireless open-ear headphones and motion tracking markers. The room contains six physical models of six different musical instruments, each with a motion tracking marker array, each corresponding to an individual track of a song. Virtual sound sources for each track are individually spatialized to the location of the corresponding instrument. This allows the participants to pick up a given instrument – which triggers the playback of the associated track—and place it wherever they like in the capture space, with the audio for that instrument appearing to emanate from the correct location. Once each instrument is placed, the participants may then freely move

through the complete sound field that they have created, allowing them to, e.g., stand back and hear the recording as the musicians may have been when on stage, or to sit down next to the drummer and hear the percussion more clearly. This installation allows people to interact with music in a way that is novel and engaging, and also gives us a good testing ground for examining the plausibility and authenticity of new spatial rendering techniques and room acoustics simulations.

#### 5.2. Real/virtual sound source comparison

In this application, the room renderings described in 4.2.1 are used to present an augmented soundscape composed of real sources (loudspeakers) and binaural virtual sources that mimic the acoustic properties of other visible loudspeakers. For instance, in a musical excerpt the voice of a singer is presented over a loudspeaker, while the guitar accompanying their vocals is presented binaurally over headphones. Listeners are then asked to identify which source is being played from the headphones. Additionally, extra controls are provided to modify the level of the direct sound, early reflections and late reverberation or to fully remove the room acoustic component of the binaural renders. This allows listeners to interactively explore the perceptual importance of sound propagation on virtual audio for augmented reality and its importance on the perceived realism.

### 6. Conclusion

Augmented Reality devices will provide a novel framework with which users can interact with the world and each other. Understanding these interactions is a high dimensional problem involving many sensory modalities and requires novel solutions to gain insight. The Interactive Auralization Plat-

form provides a vehicle to gather data about how users with augmented abilities may interact with their devices, environments, and each other by integrating currently available technology in an experientially meaningful way. The use of audio, visual, and other sensing technologies as well as the ability to augment the acoustic environment of the user, allow for robust and scalable data collection that allows for the experimental evaluation of new technologies and features that would otherwise be inhibited by form factor, compute, and sensor integration challenges. The IAP has proved valuable to understand the challenges and opportunities of augmented reality technology and further evolution of the platform is planned in the future to integrate our findings and address new areas of research.

## 7. References

- [1] V. Pulkki, "Virtual sound source positioning using vector based amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, 1997.
- [2] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *J. Audio Eng. Soc.*, vol. 61, no. 1/2, pp. 17–28, 2013.
- [3] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, "A summary on acoustic room divergence and its effect on externalization of auditory events," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, June 2016.
- [4] S. V. Amengual Garí, O. Brimijoin, H. Hassager, and P. Robinson, "Flexible binaural resynthesis of room impulse responses for augmented reality research," in *EAA Spatial Audio Signal Processing Symposium*, 2019.
- [5] F. Brinkmann, A. Lindau, and S. Weinzierl, "On the authenticity of individual dynamic binaural synthesis," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 1784–1795, 2017.



# Implementation of Ambisonics Recordings in a Wave Field Synthesis System

M. Fehling, M. Nogalski, E. Wilk

*Hamburg University of Applied Sciences, Germany, Email: matthias.fehling@haw-hamburg.de*

## Abstract

The paper presents different methods of implementation for Ambisonics recordings in a wave field synthesis system. Using the Ambisonics B-format, virtual microphone signals are extracted from the recording. The signals thus generated can be placed and reproduced at the corresponding positions in the wave field synthesis system. To further improve the performance of Ambisonics recordings in the system the possibility of sound source localisation using a sound field microphone is explored. With the presented methods it is possible to determine the position of individual recorded sound sources and to create a virtual, aligned microphone signal pointing at the source. Reproducing a single sound source gives good results in the system. When playing multiple sound sources, the problems of "sweet-spot" based systems arise.

## 1. Introduction

The Hamburg University of Applied Sciences maintains the Interactive Immersive Audiolab (I<sup>2</sup>A)<sup>1</sup>, a wave field synthesis system that is used for interactive audio drama productions as well as projects by a wide variety of artists.

In order to make content production more convenient for such systems and even use existing spatial recordings, it would be helpful to be able to utilize a more common and already established audio format, such as the Ambisonics format.

Especially the flexibility of Ambisonics recordings should be made usable in the system. Thus, the production of content can be facilitated and working with the system will become more accessible.

Our WFS system was developed by FourAudio<sup>2</sup> and consists of a rectangular assembly of linear speaker arrays. There are 56 speakers on the long sides of the rectangle and 48 on the short sides resulting in a total of 208 channels with a spacing of 10 cm between each speaker. The size of the rectangle is approximately 5 m by 6 m and the room containing the system

is acoustically optimized.

The microphone used is an Ambeo microphone by Sennheiser.

Goal of the study is to play back a B-Format Ambisonics recording in the WFS system in a meaningful way.

In the following we are going to present methods used to prepare the signals for playback. This is followed by the presentation of the measurable results. The results are evaluated and an outlook for future development is given.

## 2. Extracting the Signal

The first thing that needs to be done to an Ambisonics recording is to extract a mono signal in a defined direction  $\phi$ . This can be done by applying formula (1), [7]:

$$S = W \cdot A + (1 - A) \cdot (\cos\phi \cdot X + \sin\phi \cdot Y) \quad (1)$$

where  $W$ ,  $X$ ,  $Y$  are the corresponding signals of an Ambisonics B-format:

- $W$  : omnidirectional signal

<sup>1</sup><https://i2audiolab.de/>

<sup>2</sup><http://fouraudio.com>

- $X$  : figure eight signal on the x-axis
- $Y$  : figure eight signal on the y-axis

Since the WFS system only operates in a horizontal plane, the  $Z$  signal, containing height information of the recording, is ignored.

The resulting signal  $S$  resembles the signal of a virtually directed microphone oriented to  $\phi$  degrees. The directional characteristic can be set by mixing it with the omnidirectional  $W$  signal. The factor  $A = 0\dots 1$  thus controls the characteristic of the virtual microphone. ( $A = 0$ : figure eight;  $A = 1$ : omnidirectional).

The calculated signal can already be used in the WFS system by reproducing it from a sound source at the chosen angle  $\phi$ . This way everything the microphone recorded from the direction determined by the angle  $\phi$  can be reproduced from the same direction in the WFS system.

When trying to reproduce a full  $360^\circ$  soundfield this can be done by aligning multiple sound sources in a circle around the listening area and playing back signals created with formula (1). When using this method, the recording angles of the virtual microphones should be considered. Overlapping of the recording angles should be held minimal. When using hypercardioid characteristics ( $A = 0.25$ ) for example, a setup with seven microphones with a spacing of  $51^\circ$  each would result in a minimal overlap of recording angles.

By using this method, a reproduction of the recording is possible. This way the Wave Field Synthesis is emulating a static positioning of seven loudspeakers. This would create a surround playback situation with a sweet spot and would not use the full potential of a WFS system.

It is possible to extract the spatial information of recorded sound sources from an Ambisonics recording. Thus, a recorded sound source can be located and played back at the correct position and the movement of the source can be reproduced correctly within the playback system.

### 3. Locating a Sound Source

The location of a recorded sound source can be extracted from the recording of the sound field microphone. By calculating the intensity vector  $I = [I_X \ I_Y]$  the direction of the highest intensity at any moment of the recording can be determined. The angle  $\phi$  can now be calculated from the complex argument of the intensity vector, [7]:

$$\phi = \arg(I_X + i \cdot I_Y). \tag{2}$$

In their paper "Localization of the Sound Source with the Use of the First-order Ambisonic Microphone" Wierzbicki et al. describe three methods of locating sound sources using a soundfield microphone.

- by using the RMS value,
- by using the phase information of the  $W$  signal, or

- by using the product of the sound pressure and velocity values.

Since a RMS value can not be negative, using it would only lead to values between  $0^\circ$  and  $90^\circ$ . This requires correction to  $360^\circ$ , therefore this method is not used.

Using the phase information can be problematic since the capsules of a sound field microphone still have runtime differences. Despite the compact design of the microphone, there may occur inaccuracies with this method.

Therefore, we use the third method presented by [7]. It is assumed that with the  $W$  signal the pressure component of a recorded sound source is present and the  $X$  and  $Y$  signals contain the sound velocity information of the corresponding spatial axes. Since the sound intensity is defined as the product of sound pressure and sound velocity, the intensity can be determined as follows:

$$I_X(n) = \sum_n X(n) \cdot W(n) \tag{3}$$

$$I_Y(n) = \sum_n Y(n) \cdot W(n) \tag{4}$$

with  $(n)$  representing the sample number.

This method results in angles between  $0^\circ$  and  $180^\circ$ , which makes the correction to  $360^\circ$  much easier.

To localize a sound source from a recording, we use formula (3) and (4) to determine  $I_X$  and  $I_Y$ . With these values the intensity vector  $I = [I_X \ I_Y]$  is set up and with formula (2) the angle of the vector is determined. This angle corresponds to the direction of the highest intensity, which matches with the direction of the sound source.

The angles calculated as described are transferred to the WFS system as information for the position of the sound source. During playback, the signal created with formula (1) is placed and played back in the system at the appropriate angle. The position of the sound source corresponds to the position during recording.

### 4. Implementation

The software for WFS rendering used in the system called *wonder*, was developed by the TU Berlin and under GPL license. To play back a signal in the system the following information is needed: The audio signal and the location that signal should be played from.

Using formula (1) the signal can be produced at the desired angle. The localisation process described in section (3) is applied to reproduce the movement of the recorded sound source. This results in a list of  $\phi$  values for the duration of the signal. By using OSC (Open Sound Control) messages the location of the sound source is communicated to the system. Since only angles can be determined the distance to the center of the system was set to 6 m, just outside the radius of the physical loudspeaker array, to avoid problems that can

arise when placing sources inside the physical speaker setup. Synchronicity between the signal and its location is achieved by starting the playback and the movement at the same time.

Several possible use cases were tested.

Firstly a single sound source at a known angle was recorded. The mono signal was generated using formula (1) and the signal is played back from the corresponding position in the WFS system.

Secondly multiple sound sources are created in a full circle in an attempt to generate a full 360° soundscape. The angles of the sources are chosen to keep the overlapping of the recording angles at a minimum.

In order to check the accuracy of the method described in section 3, a single moving sound source is recorded with the Ambeo microphone while it is simultaneously being tracked by an infrared tracking system. Thus a deviation from the actual position can be determined.

From the recording, angle  $\phi$  is determined using formulas (3) and (2) for every 10 ms of the signal. The time interval of 10 ms corresponds to a spatial resolution of approximately 3.34 cm which is suitable for the speed at which the sound source was moved. The faster the sound source is moving, the shorter the chosen time interval should be.

For every angle  $\phi$ , the audio signal is generated using formula (1). As microphone characteristic we chose a hypercardioid ( $A = 0.25$ ) in order to focus the signal on the sound source. This results in a number snippets of 10 ms length which are composed together to the final audio signal of the moving source.

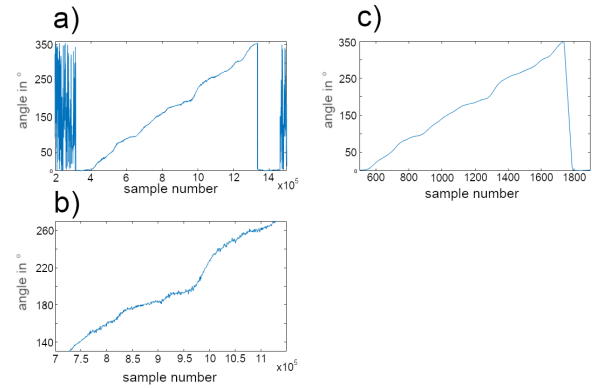
The implementations and their effects on listening impressions were evaluated by listening tests.

## 5. Results

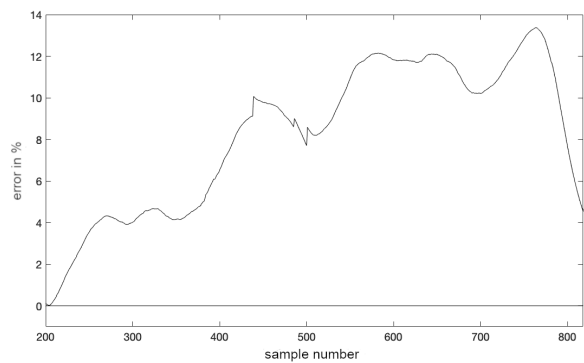
When playing back a single sound source, the audio signal only contains the information from the desired angle. This is clearly audible in the WFS system when comparing the signal reproduced with a correct  $\phi$  value to a signal that was aligned at a different direction.

This method of playback is suitable when playing back the recording of a single sound source. It is not possible to reproduce movement via this method.

When using multiple sources for playback (each source with its individual audio signal) the movement of the sound source can be heard in the WFS system. With this method the system is simulating a static loudspeaker positioning. When the recorded sound source is moving through an area without a speaker placed, the setup creates a phantom source, similar to conventional stereo and surround setups. The movement can be heard and followed by the listener, but only in a confined area in the middle of the system. If the listening position lies outside of this "sweet spot", locating the signal becomes difficult and the audio signal jumps between the virtual speakers. It is possible to implement Ambisonics



**Fig. 1:** (a): Calculated angles, (b): enlargement of (a), (c): calculated values after filtering



**Fig. 2:** Percentage error between calculated and tracked values

recordings in this way and it might be suitable to use them for ambience recordings. Though it should be avoided to generate a sweet spot inside a WFS system, as it is a benefit of a WFS system to not have a "sweet spot".

The third method consists in using the location data from the recording to position the sound source correctly in the WFS system.

Fig. 1 shows the calculated values for  $\phi$ . During recording the source was moved around the microphone in a full circle. Therefore values between 0° and 360° are expected. As Fig. 1 a) shows these values can be extracted from the recording. When playing back the signal using these values for  $\phi$  the sound source is not following a clear path, because the position of the source in the WFS system is changing rapidly. These changes in the calculated angles can be seen in fig. 1 b) and lead to audible artifacts during playback. The calculated values are therefore smoothed using a low pass filter. Fig. 1 c) shows the  $\phi$  values after smoothing. Playing back the signal using smoothed angle values yields a much better result. The movement of the source can be followed from any position inside the WFS system.

Fig. 2 shows the percentage error between calculated angle  $\phi$  and the values of the infrared tracking system. Comparing these values results in a maximum error of about 13 %, with a significant increase in the error at low signal amplitude.

The filterlength should be chosen as small as possible while

still being long enough to smooth the data to avoid large jumps in the position.

## 6. Summary

We have shown that it is possible and useful to play back Ambisonics recordings in a Wave Field Synthesis system.

Single audio signals can be extracted from the recording and played back at the desired position. This method, though possible, is not recommended for a WFS system since a soundfield microphone is not required. A simple recording with a conventional microphone in a dry environment would yield the same, if not better results.

Distributing multiple signals at their corresponding positions in a circle leads to a spatial play back utilizing phantom sources. This method is not recommended, since it shows the same restrictions of conventional surround setups: a "sweet spot" resulting in a confined listening area.

Most promising is the approach of locating the recorded sound source and transferring the location data to the WFS system. Locating the sound source with a soundfield microphone is possible with a small error. When playing back the signal while controlling the movement accordingly, the result is a playback situation where the source can be located at the correct position. Using the system like this, the advantages of WFS systems consisting of a good spatial localisation inside the system at any point, are utilized.

Nevertheless, the method to localize the sound source presented here has its restrictions. Since only the direction of the source can be calculated, it is not possible to determine the distance between the source and the microphone. This could be done by using two microphones and calculating the intensity vector for each microphone, therefore making it possible to determine the intersection of those vectors.

Another restriction is the possibility to process only a single sound source. If the recording contains two or more sources, the resulting intensity vector does not necessarily point to the correct source. This would be a task for future development, since being able to detect multiple sound sources from a single recording and placing them correctly within the WFS system would enhance the effect in the system.

## 7. References

- [1] Bates, Enda and Dooney, Sean and Gorzel, Marcin and O'Dwyer, Hugh and Ferguson, Luke and Boland, Francis M. : Comparing Ambisonic Microphones—Part 2. Audio Engineering Society Convention 142, 2017 May
- [2] Benjamin, Eric and Chen, Thomas: The Native B-Format Microphone. Audio Engineering Society Convention 119, 2005
- [3] Fohl, Wolfgang: Sound-Perception-Performance. 243–255, Springer, 2013
- [4] Fohl, Wolfgang and Wilk, Eva: Enhancements to a Wave Field Synthesis System to Create an Interactive Immersive Audio Environment. Proc. 3rd Int. Conf. on Spatial Audio, 2015.
- [5] Frank, Matthias and Zotter, Franz and Sontacchi, Alois: Producing 3D Audio in Ambisonics. Proceedings of the AES International Conference, 2015
- [6] Pulkki, Ville: Spatial Sound Reproduction with Directional Audio Coding. J. Audio Eng. Soc. Volume 55, 2007
- [7] Wierzbicki, J. and Malecki, P. and Wiciak, J.: Localization of the Sound Source with the Use of the First-order Ambisonic Microphone. Acta Physica Polonica A, Vol. 123 (2013) DOI: 10.12693/APhysPolA.123.1114
- [8] Woszczyk et al.: Tetrahedral Microphone: A Versatile "Spot" and Ambience Receiver for 3D Mixing and Sound Design, 2018



## Full Reviewed Paper at ICSA 2019

Presented\* by VDT.

### The Distribution of Ambisonic and Point Source Rendering to Ethernet AVB Speakers

S. Devonport<sup>1</sup>

R. Foss<sup>2</sup>

<sup>1</sup> Rhodes University, South Africa, Email: [tonetechnician@gmail.com](mailto:tonetechnician@gmail.com)

<sup>2</sup> Rhodes University, South Africa, Email: [r.foss@ru.ac.za](mailto:r.foss@ru.ac.za)

#### Abstract

Point source rendering is used by many object-based audio systems to mix audio objects to loudspeaker arrangements. Algorithms such as Distance-Based Amplitude Panning and Vector-Base Amplitude Panning allow for audio objects to have their locations rendered with high precision. It has been shown that in the context of loudspeaker rendering, point sources rendered with Ambisonics are often spatially blurred. However, Ambisonics does have the advantage of being able to create interesting spatial audio effects and ambient scenes can be recorded using Ambisonic microphones. This paper intends to highlight the advantages that may be gained by combining Ambisonics with virtual point source rendering. It is well known that the processing required for rendering both point source and Ambisonics can have a large overhead. To mitigate this, a distributed spatial audio system based on Ethernet AVB and distributed endpoint processors is modified to incorporate both point source rendering and Ambisonics. An example is given of how point source rendering can be integrated with Ambisonics using this system with existing software.

#### 1. Introduction

3D immersive audio can be described as the process of creating and rendering spatial audio content to a loudspeaker arrangement or headphones. With the advent of consumer VR/AR systems, there is a need for new tools that are able to render both accurate audio objects and spatially realistic ambience. Ambisonics has become a defacto standard for VR/AR productions and content is now being widely released in Ambisonic format designed mostly for headphone listening. When using Ambisonics in headphone listening, point sources are able to be rendered with high precision however there are some aspects of headphone listening that may cause breaks in immersion due to the fact that:

1. Headphones by design are unable to reproduce subsonic frequencies that enhance immersion.
2. Headphone based listening is exclusive to the person wearing the headphones.

One problem that could prevent Ambisonics from being more accepted in consumer multichannel speaker systems is the spatial precision lost due to spatial blur induced by the limitation of the loudspeaker configuration and the listening environment [1]. As such, when rendering precise virtual point sources to loudspeakers, it would be preferable to use a virtual point source rendering algorithm such as distance-based amplitude panning (DBAP) or vector-base amplitude panning (VBAP) [2] [3]. Whilst Ambisonics can cause spatial blur of encoded virtual point sources, this can be less problematic for Ambisonic recording and spatial ambience effects. It seems appropriate then to use point source panning in tandem with Ambisonics audio.

Point source panning using VBAP allows track objects to be localized precisely in loudspeaker layouts using loudspeaker triplets. These track objects can be fed signals that would normally be fed to actual speakers and these objects can now be considered 'virtual loudspeakers'. This concept has already

been used in the All Round Ambisonic Decoding (AllRAD) technique which uses VBAP generated point sources to generate an ideal loudspeaker layout for Ambisonics decoding, however, it can also be applied to surround sound speaker layout remapping [4]. This allows standard surround sound content to be played back alongside Ambisonics and object-based audio, on irregular speaker environments. Not only does this create interesting creative possibilities, but it also allows for backwards compatibility with channel-based surround sound content that is currently available in many movies and games.

The ImmerGo spatial audio workstation provides a framework to implement such features [5]. It already incorporates object-based point source panning with DBAP and VBAP, and can be modified to incorporate Ambisonic rendering as it uses object metadata to describe loudspeaker positions. Its client-server based distributed architecture decouples its usage from particular sound source software such as a digital audio workstation (DAW), and this enables it to be easily integrated into any audio project's workflow. The use of Ethernet AVB and distributed network processors mitigates the processing demands of these different rendering algorithms and makes it scalable to any number of loudspeakers [6].

The research in this paper modifies ImmerGo to include both distributed virtual point source rendering and Ambisonics rendering. The VBAP algorithm is utilized to provide speaker remapping that allows for playback of surround sound content on irregular loudspeaker layouts and the AllRAD algorithm is used to decode up to 4<sup>th</sup> order Ambisonics to irregularly spaced loudspeakers.

This paper will continue with an overview of channel-based audio, object-based audio and Ambisonics, highlighting key benefits of each. This will be followed by an explanation of how these immersive approaches were merged within ImmerGo to create a system incorporating their combined benefits. Finally, there will be a description of an installation that utilizes this system.

## 2. Representations of 3D Immersive Audio

Channel-based audio (CBA), object-based audio (OBA) and scene-based audio (SBA) are representations used in current state-of-the-art immersive audio rendering systems. These have been developed to simplify the process of creating and distributing spatial audio content correctly to different loudspeaker arrangements. Each of these representations lends themselves to different rendering procedures that are required to generate and feed audio sources to loudspeakers. Currently, there are a few systems available that have been developed to render these different representations alongside each other, notably the MPEG-H Renderer, the European Broadcast Union (EBU) ADM Renderer (EAR) which is now being implemented in the ITU-R BS.2127 [7] [8] [9] [10].

### 2.1. Channel-Based Audio

CBA can be considered a 'loudspeaker first' spatial audio technique. A content creator will mix composition into a multichannel wav file. It assumes the loudspeaker arrangement used for playback will be the same loudspeaker arrangement that was used when generating the mix. This has led to the well-known ITU surround sound loudspeaker arrangements used in home theatre [11]. This format requires a relatively simple rendering procedure, as there is a one to one mapping between audio channel and loudspeaker.

While the CBA format has worked relatively well in consumer audio distribution, it has avoided dealing with the problem of non-standard loudspeaker configurations: if you play a CBA mix to an irregular layout of loudspeakers, there is a good chance it will not sound the way the content creator intended. There are solutions available that mitigate this such as the MPEG-H renderer which provides tools that can translate between different CBA layouts [12].

### 2.2. Object-Based Audio

Object-oriented distribution formats have been developed to allow for spatial audio playback to be compatible with any loudspeaker arrangement by using powerful processors that render the spatial audio content at playback time. The object-based format incorporates audio sources, the positional information of these sources and the loudspeaker playback environment. This has been aptly named OBA. OBA represents each audio channel associated with location, spread and directionality metadata. When creating object-based audio content, the metadata allows the spatial scene to be rendered correctly on non-standard and standard loudspeaker arrangements alike [13]. At the playback stage, the metadata associated with an audio channel is fed into an algorithm that generates the loudspeaker feeds so that the audio channel is correctly positioned in 3D space.

OBA has been incorporated in systems such as Dolby Atmos, DTS:X and Auro3D. It comprises a significant part of the MPEG-H 3D Audio standard. The EBU has released an open source Audio Definition Model (ADM) metadata format that has been a resource for the ITU Audio Definition Model [14] [15] and is used by the open source EAR renderer and the ITU-R BS.1770 [16].

There are a variety of object-based rendering algorithms that can be used to render audio object positions according to their metadata, most notably VBAP and DBAP. While both these algorithms render audio objects using metadata there are differences in how they render the audio in relation to the listening position. VBAP assumes there to be a listener centred at an origin point or sweet spot, whilst DBAP does not assume this.

### 2.3. Scene-Based Audio and Ambisonics

SBA could be considered a combination of CBA and OBA [9]. It encodes an infinite number of audio objects into a known set of audio channels known as the *audio scene*. This encoded format is loudspeaker agnostic and the format must be decoded at the loudspeaker endpoint to be heard correctly



[17]. There are different formats that describe the Ambisonic soundfield, however, it has become a standard to use the AmbiX format which is output by various plugins and is the format used in the research in this paper [18] [19].

Ambisonics uses a set of encoding functions based on the spherical harmonic transform functions [20]. These functions are used to encode the positions of audio objects that lie on a sphere into a finite set of audio channels called the Ambisonic soundfield. An Ambisonic soundfield can be captured using Ambisonic microphones, or synthesized by multiplying a mono channel by each encoder function to form the Ambisonic encoded audio channels.

The number of channels in the encoded signal is proportional to the order according to:

$$\text{channels} = (\text{order} + 1)^2$$

1<sup>st</sup> order content is sometimes referred to as B-format and has four channels, and higher order content has more, with higher order content having a higher spatial resolution [21].

An in-depth discussion into the decoder formulation is omitted as it has already been covered in depth in numerous other publications [22]. However, it is important to know that at the decoding stage, these audio channels are summed together and fed to loudspeakers according to decoder scaling values that are calculated according to the loudspeaker position. These decoder values ensure that all the loudspeakers are playing audio from the Ambisonic signal, with the sound scene directionality being induced by the weighting and phase inversions of each audio signal from the encoded Ambisonic soundfield. Since all the loudspeakers are playing at once, the decoding can cause some spatial blur when listening to Ambisonic content on loudspeaker arrays with insufficient, incorrectly spaced loudspeakers, or when listening at an off-centre position [23].

Due to the geometrical nature of Ambisonics, the encoded channels are also able to be manipulated and transformed efficiently in real-time using processing matrices. This is particularly useful in headtracked environments for headphone reproduction of VR audio. This feature has also allowed for interesting and efficient Ambisonic effects to be created [24]. These effects are easily incorporated into workflows currently used for media production [25]. As well as this, there are a variety of Ambisonic based processors that have been developed for both game audio engines<sup>1</sup> and digital audio workstations<sup>2,3</sup>.

In the last few years, many Ambisonic microphone arrays have been released that can be used to record spatial audio that can be played alongside VR games and video. The majority of these microphones are in a tetrahedral arrangement that records in A-format, which is converted to the B-format used

in Ambisonics. These microphones capture ambience and directionality sufficiently accurately, however higher order microphones capture sound field directionality more accurately. There is also the well-known Eigenmike<sup>4</sup> which is a 32-channel microphone array that is able to generate higher order Ambisonic files up to 4<sup>th</sup> order. Some free-to-download Ambisonic recordings made with the Eigenmike known as the Eigenscape can be found online [26].

Of considerable interest is that these microphones are able to capture a 3D directional impulse response of an environment that can be used as a convolution filter for other Ambisonic encoded content. These so-called Directional Room Impulse Responses (DRIRs) provide accurate 3D modelling of acoustic spaces [27]. Recently, a few databases of DRIRs have been converted into the Spatial Oriented Format for Acoustics (SOFA) convention [28] [29]. SOFA provides a standardized format which can be used interchangeably between different systems. This provides a promising basis for the growth of these applications in the future.

### 3. ImmerGo Spatial Audio Workstation

The ImmerGo spatial audio workstation provides a client-server-based approach to immersive audio rendering and is built on web technologies [5]. It allows a user to render the location of multiple virtual point sources in an environment using any device with a browser. ImmerGo's approach to immersive audio rendering employs a distributed processing model that moves the final audio rendering processing out to multiple endpoint processors attached to loudspeakers. Each processor is dedicated to the loudspeaker it is attached to. This ensures a scalable solution, as any loudspeaker added to the loudspeaker array also incorporates the additional processing power.

As shown in **Fig. 1**, a user is able to select a track and control its position, level and spread angle within the loudspeaker array. ImmerGo also can control a DAW transport using an internal MIDI bus. As well as this, track object metadata parameters are able to be recorded and automated according to the internal clock or MIDI timecode from an external source. When playing this automation, the ImmerGo track object model is updated according to MIDI time code quarter frames at roughly 8ms intervals.

#### 3.1. ImmerGo Track and Room Object Model

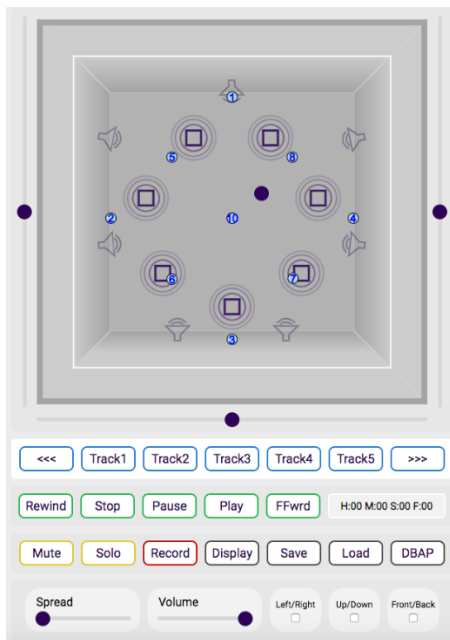
Shown in **Table 1** is ImmerGo's track object model. This model contains parameters which are used by the renderer to generate mixing values used for the endpoint processors that mix the spatial composition correctly. The dynamic parameters are able to be changed in real-time using the ImmerGo UI. The track ID does not change.

<sup>1</sup>  
[https://www.audiokinetic.com/library/edge/?source=Help&id=using\\_ambisonics\\_in\\_plugins](https://www.audiokinetic.com/library/edge/?source=Help&id=using_ambisonics_in_plugins)

<sup>2</sup> IEM plugins - <https://plugins.iem.at/>

<sup>2</sup> SPARTA/COMPASS plugins - [http://research.spa.aalto.fi/projects/sparta\\_vsts/plugins.html](http://research.spa.aalto.fi/projects/sparta_vsts/plugins.html)

<sup>4</sup> Eigenmike - <https://mhacoustics.com/products>



**Fig. 1:** The ImmerGo Client.

Track Object Model	
- Position (x,y,z)	Dynamic
- Spread	Dynamic
- Volume	Dynamic
- Track ID	Static
- Mute	Dynamic
- Solo	Dynamic
- Selected	Dynamic

**Table 1:** The ImmerGo Track Object Model

The loudspeaker object model shown in **Table 2** is used by the spatial audio renderer to calculate the correct loudspeaker signals for a particular audio channel. Loudspeakers are able to be positioned using the ImmerGo UI which generates the loudspeaker object metadata. Each audio channel fed to the loudspeaker is able to be scaled and delayed. In this research, only the scaling function is used.

Loudspeaker Model	
- Position (x,y,z)	
- Mix values for each audio channel	
- Delay levels for each audio channel	

**Table 2:** The ImmerGo Loudspeaker Model

### 3.2. Ethernet AVB

The loudspeaker processors interface to an Ethernet AVB network. The Ethernet AVB standard provides the framework to allow for a synchronous real-time audio network. It adds two standards on top of three older standards to enable the appropriate quality of service for real-time audio delivery and control. These are the 1722 Audio Video Transport protocol (AVTP) and the 1722.1 Audio Video Discovery, Enumeration, Control and Connection Management protocol (AVDECC) [30] [31].

These provide two important benefits that are used extensively by the ImmerGo system:

1. AVTP provides the ability to stream multichannel audio synchronously to multiple loudspeaker processors distributed on a network.
2. AVDECC provides the ability to control multiple loudspeaker processors distributed on a network in real-time using the AVDECC Enumeration and Control Protocol (AECp).

### 3.3. Distributed Endpoint Processors

Each distributed loudspeaker processor used by ImmerGo contains an XMOS microcontroller and SHARC DSP chip [32] [33]. This provides each loudspeaker processor with the capability to have audio mixed in real-time according to the output of various spatial audio rendering algorithms housed within the ImmerGo server. Furthermore, the processors and speakers are able to be powered over Ethernet using the PoE+ protocol.

A core component of the endpoint processor is its multi-in multi-out (MIMO) mixer matrix that is controlled using the AECp protocol. The MIMO mixer can scale and phase invert each of the 32 channels of audio coming from the Ethernet AVB stream before summing them to either two loudspeakers attached to the processor. The values used to scale the output are generated by the ImmerGo server according to object metadata.

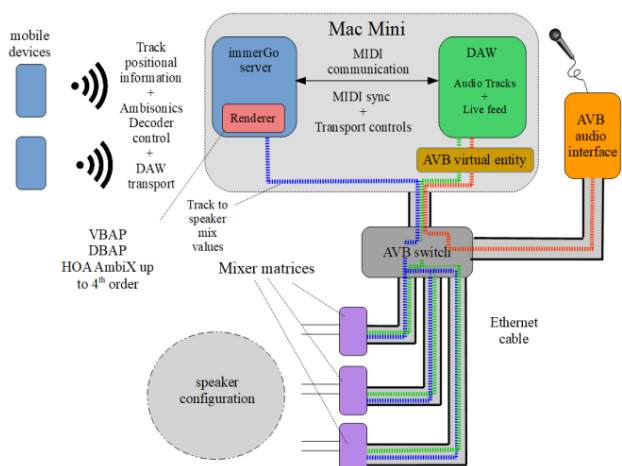
### 4. Modifications to ImmerGo

As shown in **Fig. 2** below, ImmerGo’s UI and spatial audio renderer is modified to include the capability to control and render higher order Ambisonics alongside VBAP and DBAP. This allows for a variety of Ambisonics recordings and effect chains to be played alongside virtual point source rendered content. Other AVB interfaces with live feeds are also able to be included in the network using the native Apple AVB virtual entity.

We see in orange the live audio feed from a microphone passing through the network to the DAW. Within the DAW, the live feed can be processed and streamed out alongside other audio content housed within the DAW shown in green. When a user interacts with a track object, Ambisonics decoder or speaker remapping function on the UI, the updates are sent over a web socket which is then parsed within the server. The server’s renderer then uses these parameters to render mixer values for the loudspeaker processor mixer matrices which are sent to the endpoint processors using AVDECC AECp messages.

There were three main considerations taken into account when implementing these modifications:

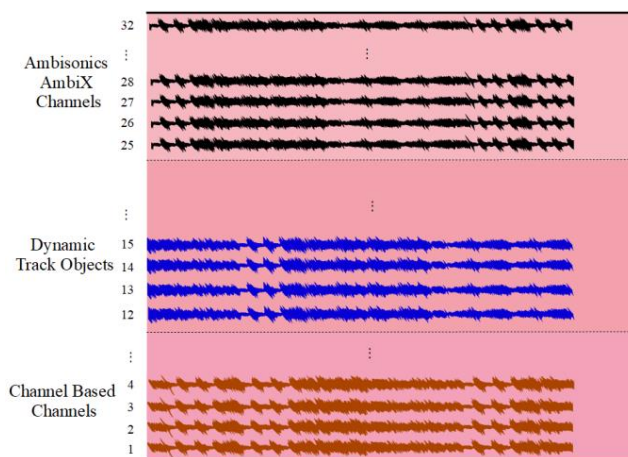
1. The mixer matrix limit of 32 channels.
2. Changes to listening position across different speaker environments.
3. Speaker remapping using point source rendering.



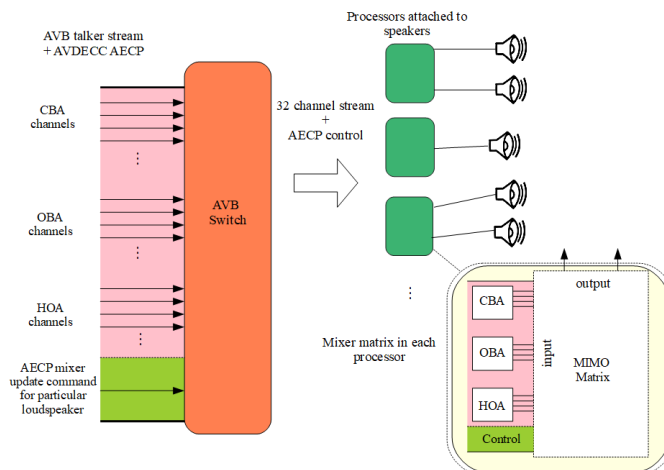
**Fig. 2:** Modifications to the ImmerGo system to include Ambisonics decoding and speaker remapping control.

### 4.1. Audio Transport Allocation to the Mixer Inputs

As shown in Fig. 3, the audio transport bus is split into three sub-busses that are fed the appropriate audio channels pertaining to the style of rendering being performed. The audio channel allocation is changed dynamically according to the number of channels required for each rendering algorithm. Because of this, there is a channel allocation trade-off. The higher the order of Ambisonic decoding, the less available channels there are for the point source rendering that is used for OBA and CBA content. This is realised in how a user can interact with the UI. If a part of the audio transport is dedicated to Ambisonics, the user is unable to interact with that channel using the typical track object controls. However, if a part of the audio transport is dedicated to channel-based content, the user is able to fine tune the virtual speaker object position.



**Fig. 3:** Channel allocation from the audio source to the Ethernet AVB stream.



**Fig. 4:** Channel allocation of AVB stream with control message from ImmerGo to each endpoint mixer matrix.

Fig. 4 above shows how each endpoint mixer matrix input is fed according to the channel allocation given by the sub-bus components for each rendering algorithm. The channel-based audio that renders surround sound content was placed first, with object-based speaker remapping applied. Then dynamic object-based tracks were placed second and the rest of the transport was dedicated to the Ambisonic encoded signals. The mixer inputs dedicated to each format then had their mixer crosspoints updated according to the correct values pertaining to the rendering algorithm.

### 4.2. Surround Sound Speaker Remapping

As has been shown by the AllRAD approach, virtual point sources created using VBAP are able to be used as 'virtual loudspeakers' [4]. This concept is used to provide ImmerGo with the capability to playback channel-based content. By assigning a particular loudspeaker signal from the CBA content to a track object, the loudspeaker feed is able to be played from that position. In order to achieve this in practice, a central origin is needed so that each object is rendered to the correct location in the array. This origin position is known as the listening position and is calculated as the midpoint of the maximum and minimum (x,y,z) locations of the loudspeakers in the array.

DBAP would also be able to pan a virtual source, however, the algorithm is based on loudspeaker energy distribution such that all the loudspeakers are required to play the audio to keep a constant energy level. VBAP provides more precise point sources when compared to DBAP since only 3 speakers are active at once.

The option to 'remap speakers' is provided in the user interface as shown in Fig. 5. When selecting this option, controls are shown that are used to set the desired loudspeaker configuration using virtual point sources panned with VBAP, as well as control the bass management level. A selection of mono, stereo, 2.1, 4.1, 5.1 and 7.1 are currently available. The vertical offset of the listening position can be controlled using the vertical offset slider. This is used in case the virtual source positions are not in the same plane as the listener position. The

track buttons are then changed to give the user feedback about which surround channels are mapped to which location.

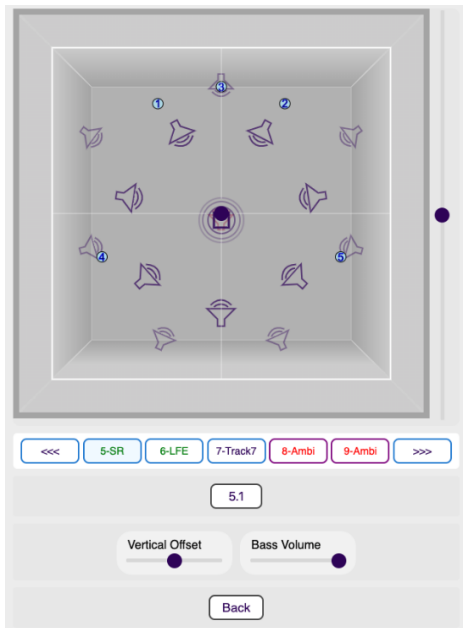


Fig. 5: Speaker remap control using ImmerGo.

### 4.3. Ambisonics Decoder Control

Traditional Ambisonics requires very strict loudspeaker layouts such as cubes or icosahedrons however in recent years, the AllRAD approach has been developed to overcome this limitation. It makes use of virtual point source panning using VBAP to create ‘virtual loudspeakers’ in the layout of T-designs that Ambisonics encoded signals are able to map to successfully [4]. This solution allows for Ambisonics to be decoded to irregular loudspeaker layouts. In order to have ImmerGo compatible with different layouts, this approach was employed.

ImmerGo was modified to incorporate the Spatial Audio Framework (SAF) <sup>5</sup> which has recently been developed at Aalto University and provides an AllRAD solution. SAF is able to generate AllRAD Ambisonics decoder values with maxRE weightings up to 7<sup>th</sup> order AmbiX for any speaker layout. On suboptimal layouts where the convex hull calculation results in large changes of gain values, an ‘imaginary’ loudspeaker is added either directly above or below the loudspeaker array to ensure a balanced energy distribution [4].

In order to use the features provided in SAF, NodeJS bindings were developed that allowed for data to be passed between the ImmerGo server’s NodeJS runtime and the SAF library. Due to the limit imposed by the endpoint mixer matrices, a maximum of 4<sup>th</sup> order (25 channels) was allowed.

When solving for the loudspeaker directions the listening position is selected as the centre of the loudspeaker array. From there, vectors are drawn to the loudspeakers and these

are used to solve for the loudspeaker directions. Each loudspeaker direction is then passed to SAF which calculates the AllRAD decoder values. These values are then sent to the endpoint mixer matrices at which point the track object control in the ImmerGo UI becomes disabled to the user. The Ambisonic decoder controls are found in a sub menu.

ImmerGo’s Ambisonic controls are shown in Fig 6. below. Loudspeaker environments may change and as such, the listening position also needs to be adjustable. If the initial point of origin is incorrect for the loudspeaker array, the sound scene can sound weighted unevenly in the vertical plane. As such, the ability to change the listening position vertical offset is provided, and is different from the vertical offset parameter of the speaker remapping option. When changing this control, the AllRAD decoders are recalculated, with the loudspeaker directions shifted according to the change in the origin. The result is the impression that the origin of the Ambisonic soundfield has shifted vertically in the loudspeaker array. Along with these controls is the option to convert from 3D normalised and a semi-normalized (N3D or SN3D) normalization scheme [4].

The Ambisonic soundfield is also able to be rotated using the relevant yaw, pitch and roll sliders. A bass management control is provided allows for changes to the volume of the omnidirectional component of the Ambisonic encoded signal feeding the subwoofers.

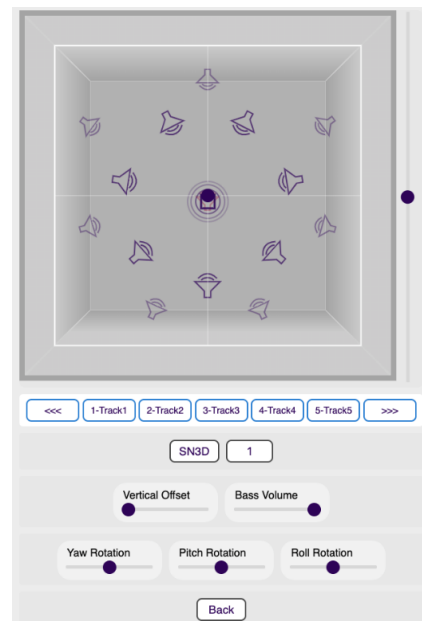


Fig. 6: Ambisonics decoder control using the ImmerGo UI.

## 5. Combining Ambisonics with Virtual Point source Rendering

To demonstrate the benefits of combining these algorithms, a system that combines the modified ImmerGo system with the Reaper DAW has been created, although any other audio software could be used. A score was built that made use of

<sup>5</sup> [https://github.com/leomccormack/Spatial\\_Audio\\_Framework](https://github.com/leomccormack/Spatial_Audio_Framework)



point source panning, speaker remapping and Ambisonics. It consisted of a male voice, Ambisonic nature recording, 5.1 sound effects, and track objects that were used to move particular sound sources within the Ambisonic scene.

The score intended to recreate more precisely the sound of a voice being inside a natural soundscape. Ambisonic recordings of a nature scene and DRIRs were captured, using the H3-VR microphone. This recording was used alongside the point source and 5.1 content. Mixed into the Ambisonic bus were other audio effects generated by Ambisonic plugins to create spatial echoes and virtual source position modulations, some of which were convolved with Ambisonic DRIRs taken in different environments.

The channel allocation for the content in this score is:

1. **Channels 1-6:** Previously created 5.1 sound effects including close miked natural sounds such as leaves rustling and animal sounds.
2. **Channels 7-16:** Track with different spot mikes of birds and bug sounds.
3. **Channels 29-32:** 1<sup>st</sup> order AmbiX recordings from the H3-VR microphone mixed with other 1<sup>st</sup> order Ambisonic spatial effects of birds.

The project was run as follows:

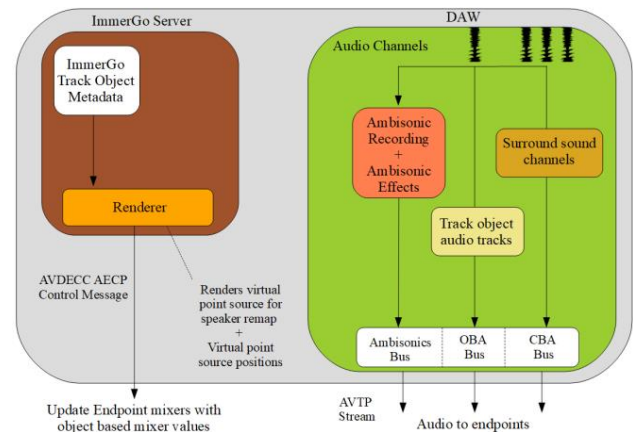
1. At startup, the speaker configuration was input to the ImmerGo system using the mobile device GUI.
2. These positions were used by the SAF to calculate and update the endpoint mixers for the Ambisonic decoders according to the loudspeaker positions
3. The speaker remapping was set to use 5.1 rendering.
4. The Ambisonic decoder was adjusted according to the content's format.
5. The audio content was played from the DAW and track objects were able to be controlled alongside the Ambisonic and surround sound content.

This system highlights some key benefits when combining these various techniques to create an immersive audio soundscape. The following capabilities are provided:

- Virtual point source panning using DBAP and VBAP.
- 5.1 surround sound content playback.
- 3D Ambisonic recordings for the rendering of spatial audio recordings.
- Ambisonics effect processing for efficient spatial audio effects.

**Fig. 7** below shows how these were combined using ImmerGo along with the Reaper DAW. The Ambisonic decoder was able to decode the Ambisonics recordings and effects that were summed into the Ambisonics transport bus. As well as this, the speaker remapping option allowed for the channel-based content to be rendered on the irregular layout.

Furthermore, individual track objects were able to be moved around the speaker array according to automation tracks.



**Fig. 7:** Object metadata controlling the ImmerGo point source renderer alongside Ambisonics rendering.

## 6. Conclusion

This paper has covered the various representations of immersive audio with their associated rendering algorithms. This information was used to modify the ImmerGo spatial audio workstation so that it is able to perform rendering for channel-based, object-based and scene-based audio alongside each other. In particular, sub menus were added that provide the necessary controls for speaker remapping and Ambisonic decoder setup. A possible project layout that incorporates the rendering of channel-based, object-based and Ambisonic content alongside one another is given. This project highlights how a rich combination of channel-based audio, point source panning and Ambisonics effects can be rendered simultaneously.

## 7. References

- [1] G. Marentakis, F. Zotter and M. Frank, "Vector-Base and Ambisonic Amplitude Panning: A Comparison Using Pop, Classical, and Contemporary Spatial Music," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 945-955, 2014.
- [2] T. Lossius, P. Baltazar and T. Hogue, DBAP--distance-based amplitude panning, Ann Arbor, MI: Michigan Publishing, University of Michigan Library, 2009.
- [3] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, vol. 45, pp. 456-466, 1997.
- [4] F. Zotter and M. Frank, "All-round ambisonic panning and decoding," *Journal of the Audio Engineering Society*, vol. 60, pp. 807-820, 2012.
- [5] R. Foss and A. Rouget, "Immersive Audio Content Creation Using Mobile Devices and Ethernet AVB," in *Audio Engineering Society Convention 139*, 2015.
- [6] S. Devonport and R. Foss, "An Investigation into the Distribution of 3D Immersive Audio Renderer Processing to Speaker Endpoint Processors," in *VDT Tonmeistertagung*, Cologne, 2018.

- [7] A. Murtaza, J. Herre, J. Paulus, L. Terentiv, H. Fuchs and S. Disch, "ISO/MPEG-H 3D Audio: SAOC 3D Decoding and Rendering," in *Audio Engineering Society 139*, New York, 2015.
- [8] European Broadcast Union, "Tech 3388: ADM Renderer for Use in Next Generation Audio Broadcasting," EBU, Geneva, 2018.
- [9] S. Shivappa, M. Morrell, D. Sen, N. Peters and S. M. A. Salehin, "Efficient, Compelling, and Immersive VR Audio Experience Using Scene Based Audio/Higher Order Ambisonics," in *AES International Conference on Audio for Virtual and Augmented Reality*, Los Angeles, 2016.
- [10] International Telecommunications Union, "ITU BS.2127: Audio Definition Model renderer for advanced sound systems," 2019.
- [11] International Telecommunications Union (ITU), "ITU-Report BS.2159-6," International Telecommunications Union (ITU), Geneva, 2015.
- [12] J. Herre, J. Hilpert, A. Kuntz and J. Plogsties, "MPEG-H 3D Audio—The New Standard for Coding of Immersive Spatial Audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770-779, 2015.
- [13] R. Bleidt, A. Borsum, H. Fuchs and S. M. Weiss, "Object-Based Audio: Opportunities for Improved Listening Experience and Increased Listener Involvement," 2014.
- [14] European Broadcasting Union, "TECH 3364 Audio Definition Model Metadata v2.0," European Broadcasting Union, Geneva, 2018.
- [15] International Telecommunications Union, "Recommendation ITU-R BS.2076-1," International Telecommunications Union, Geneva, 2017.
- [16] International Telecommunications Union, "ITU-R BS.1770: Algorithms to measure audio programme loudness and true-peak audio level," 2015.
- [17] N. Peters, D. Sen, M.-Y. Kim, O. Wuebbolt and S. M. Weiss, "Scene-based Audio Implemented with Higher Order Ambisonics," *SMPTE Motion Imaging Journal*, vol. 125, no. 9, pp. 16 - 24, 2016.
- [18] C. Nachbar, F. Zotter, E. Deflelie and A. Sontacchi, "AmbiX - A Suggested Ambisonics Format," Lexington, 2011.
- [19] L. McCormack and A. Politis, "SPARTA & COMPASS: Real-Time Implementations of Linear and Parametric Spatial Audio Reproduction and Processing Methods," in *AES International Conference on Immersive and Interactive Audio*, York, 2019.
- [20] M. A. Gerzon, "Periphony: With-height Sound Reproduction," *Journal of the Audio Engineering Society*, vol. 21, no. 1, pp. 2-10, 1973.
- [21] S. Bertet, J. Daniel, L. Gros, E. Parizet and O. Warusfel, "Investigation of the Perceived Spatial Resolution of Higher Order Ambisonics Sound Fields: A Subjective Evaluation Involving Virtual and Real 3D Microphones," in *AES 30th International Conference: Intelligent Audio Environments*, Finland, 2007.
- [22] D. Artega, "Lecture Notes: Introduction to Ambisonics," in *Audio 3D – Grau en Enginyeria de Sistemes Audiovisuals*, Universitat Pompeu Fabra, 2015.
- [23] L. S. R. Simon, H. Wuethrich and N. Dillier, "Comparison of Higher-Order Ambisonics, Vector- and Distance-Based Amplitude Panning using a hearing device beamformer," in *4th International Conference on Spatial Audio*, Graz, 2017.
- [24] D. Rudrich, F. Zotter and M. Frank, "Efficient Spatial Ambisonic Effects for Live Audio," in *29th Tonmeisteragung - VDT International Convention*, Cologne, 2016.
- [25] F. Zotter and M. Frank, "Signal Flow and Effects in Ambisonic Productions," in *Ambisonics - A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, Graz, Springer Open, 2019, pp. 99 - 130.
- [26] M. C. Green and D. Murphy, Composers, *Eigenscape* - <https://zenodo.org/record/1012809>. [Sound Recording]. University of York. 2017.
- [27] A. Rahman, "Portable Ambisonic Impulse Response System (P.A.I.R.S)," March 2017. [Online]. Available: <https://cnmat.berkeley.edu/projects/pairs>. [Accessed 17 June 2019].
- [28] A. Pérez-López and J. De Muynke, "Ambisonics Directional Room Impulse Response as a New Convention of the Spatially Oriented Format for Acoustics," in *144th AES Convention*, Milan, 2018.
- [29] Audio Engineering Society, "AES69: AES standard for file exchange - Spatial acoustic data file format," 2015.
- [30] IEEE, *Std 1722 (IEEE Standard for a Transport Protocol for Time-Sensitive Applications in Bridged Local Area Networks)*.
- [31] IEEE, *Std 1722.1 (IEEE Standard for Device Discovery, Connection Management, and Control Protocol for IEEE 1722 Based Devices)*.
- [32] XMOS, "XMOS Microcontrollers," [Online]. Available: <https://www.xmos.com/developer/products/silicon>. [Accessed 22 June 2019].
- [33] SHARC, "SHARC ADSP21489," [Online]. Available: <http://www.analog.com/en/products/audio-video/audio-signal-processors/sharc/adsp-21489.html>. [Accessed 22 June 2019].
- [34] "IEEE Standard for Local and metropolitan area networks--Bridges and Bridged Networks," *IEEE Std 802.1Q-2014 (Revision of IEEE Std 802.1Q-2011)*, pp. 1-1832, 12 2014.
- [35] "IEEE Standard for Local and metropolitan area networks--Audio Video Bridging (AVB) Systems," *IEEE Std 802.1BA-2011*, pp. 1-45, 9 2011.
- [36] "IEEE Standard for Local and Metropolitan Area Networks - Timing and Synchronization for Time-Sensitive Applications in Bridged Local Area Networks," *IEEE Std 802.1AS-2011*, pp. 1-292, 3 2011.
- [37] D. Kostadinov, J. D. Reiss and V. Mladenov, "Evaluation of distance based amplitude panning for spatial audio," in *ICASSP*, 2010.



## Full Reviewed Paper at ICSA 2019

Presented \* by VDT.

### How positioning inaccuracies influence the spatial upsampling of sparse head-related transfer function sets

Christoph Pörschmann<sup>1,†</sup>, Johannes M. Arend<sup>1,2</sup>

<sup>1</sup> *Institute of Communications Engineering, Technische Hochschule Köln, D-50679 Cologne, Germany*

<sup>2</sup> *Audio Communication Group, TU Berlin, D-10587 Berlin, Germany*

<sup>†</sup> *Corresponding author, E-mail: Christoph.Poerschmann@th-koeln.de*

#### Abstract

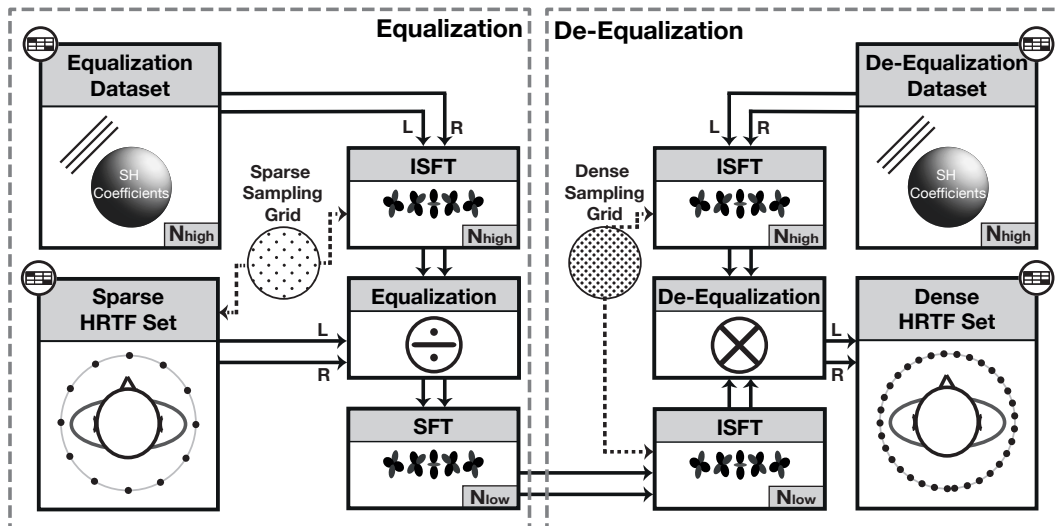
Determining full-spherical individual sets of head-related transfer functions (HRTFs) based on sparse measurements is a prerequisite for various applications in virtual acoustics. To obtain dense sets from sparse measurements, spatial upsampling of sparse HRTF sets in the spatially continuous spherical harmonics (SH) domain can be performed by an inverse SH transform. However, this involves artifacts caused by spatial aliasing and order truncation. In a previous publication we presented the SUPDEq method (Spatial Upsampling by Directional Equalization), which reduces these artifacts by a directional equalization prior to the SH transform. Generally, apart from the spatial resolution of the HRTF set, measurement inaccuracies, for example caused by displacements of the head during the measurement, can influence the spatial upsampling as well. By this direction-depending temporal and spectral deviations are added to the dataset, which in the process of spatial upsampling can cause artifacts comparable to spatial aliasing errors. To reduce the influence of the distance inaccuracies, we present a method for distance error compensation that performs an appropriate distance-shifting of the measured HRTFs. Determining the required values for the shift benefits from the directional equalization performed by SUPDEq and results in time-aligning the directionally equalized HRTFs. We analyze the influence of the angular and distance displacements on spectrum, on interaural cues and on modeled localization performance. While limited angular inaccuracies only have a low impact, already small random distance displacements cause strong impairments, which can be significantly reduced applying the proposed distance error compensation method.

#### 1. Introduction

A spatial presentation of sound sources is a fundamental element of virtual acoustic environments (VAEs). For this, monaural and binaural cues, which are mainly caused by the shape of the pinna and the head, need to be considered. In many headphone-based VAEs, head-related transfer functions (HRTFs) are applied to describe the sound incidence from a source, which is typically in the far-field, to the left and right ear incorporating both, monaural and the binaural cues.

A high number of HRTFs is required to adequately capture

these cues for all directions of incidence. Complete sets of HRTFs measured on a spherical grid can be described in the spherical harmonics (SH) domain by a decomposition into spherical base functions of different spatial orders  $N$ , where higher orders correspond to a higher spatial resolution [15, 18]. Describing sparse HRTF sets in the SH domain results in a limited order and incorporates an incomplete description of the spatial properties. This results in spatial aliasing and truncation errors. To completely consider these properties, an order  $N \geq kr$  with  $k = \omega/c$ , and  $r$  being the head radius is required [7, 14]. Performing a nearly perfect



**Fig. 1:** Block diagram of the SUPDEq method. Left panel: A sparse HRTF set is equalized on the corresponding sparse sampling grid before transformed to the SH domain with  $N = N_{low}$ . Right panel: The equalized set is de-equalized on a dense sampling grid. If required, the resulting dense HRTF set can again be transformed to the SH domain with  $N = N_{high}$ .

interpolation for frequencies up to 20 kHz leads to  $N = 32$  requiring at least 1089 measured directions when assuming  $r = 8.75$  cm and  $c = 343$  m/s.

Different studies analyzed artifacts caused by spatial upsampling of sparsely measured HRTF sets to a dense sampling grid or examined methods to reduce these artifacts (e.g. [5, 7, 9, 19]). In this context, we recently introduced the SUPDEq (Spatial Upsampling by Directional Equalization) method [12], which removes frequency-dependent ITDs and ILDs as well as elevation-dependent spectral features from the HRTFs. For this we apply spectral division (equalization) to the HRTFs with a corresponding equalization function prior to the SH transform. A directional rigid sphere transfer function (STF) can be used here as equalization function, resulting in a significantly reduced spatial order  $N$ . After spatial upsampling, a de-equalization by means of a spectral multiplication with the same equalization function is performed to recover a spatially upsampled HRTF set.

Generally speaking, the use of the proposed method is especially advantageous for measuring sets of individual HRTFs which, for example, provide a better localization accuracy in the median plane than non-individual ones [8]. However, measuring such datasets in a simple procedure with a cheap measurement setup, and under non-ideal room-acoustical conditions is a challenging task. In previous papers we already analyzed the suitability of the SUPDEq method for individual HRTF sets [13] and investigated to what extent the use of low-cost loudspeakers in reflective environments affects the HRTF measurements [11].

In this paper we analyze another critical issue. The positioning accuracy of the subject in the array during the HRTF measurement can result in distance and angular errors. These inaccuracies can on the one hand be caused by a constant shift of the listener's center position while measuring. For example, in [4] it is shown that such shifts of

the listener position significantly increase the required spatial order of the HRTF set. To compensate for this, methods for recentering the receiver by appropriate postprocessing of the measured dataset have been developed [16]. On the other hand these shifts can be non-systematic and thus be independent of the measured directions, resulting in (nearly) randomly distributed distance and angular inaccuracies. Such deviations in the measured HRTFs might be observed with sequential HRTF measurements. Furthermore, when tracking the listener's position and orientation in such a procedure, the inaccuracies of the tracking device can as well be regarded as being randomly distributed.

To investigate the influence of positioning inaccuracies on spatial upsampling, we performed a study which compares spatially upsampled sparse HRTF sets to a reference sampled on a dense grid. We analyze the influence of angular and distance errors regarding spectral differences, binaural cues and modeled localization performance. Furthermore, we investigate to what extent a method compensating the distance errors can enhance the performance. Thus the result of this study can help to obtain required boundary conditions for performing HRTF measurements.

## 2. Method

The SUPDEq method has been described and evaluated in detail in [12]. In the following we thus only briefly outline the basic concept. The corresponding block diagram is given in Fig. 1. First, the sparse HRTF set  $H_{HRTF}$  measured at  $S$  sampling points  $\Omega_s = \{(\phi_1, \theta_1), \dots, (\phi_S, \theta_S)\}$  is equalized direction-dependently with an appropriate equalization dataset  $H_{EQ}$

$$H_{HRTF, EQ}(\omega, \Omega_s) = \frac{H_{HRTF}(\omega, \Omega_s)}{H_{EQ}(\omega, \Omega_s)}. \quad (1)$$

While generally different equalization datasets can be applied, in this study a rigid sphere transfer function (STF) is used



which describes an incoming wave on a rigid sphere [18, p. 227]. The radius of the sphere corresponds to the physical dimensions of a human head and an ear position of  $\phi = \pm 90^\circ$  and  $\theta = 0^\circ$  is considered. The STF can thus be regarded as a simplified HRTF set featuring basic temporal and spectral components but without information on the shape of the outer ears or the fine structure of the head. Thus, by the equalization a time-alignment of the HRTFs is performed and direction-dependent influences of the spherical shape of the head are compensated. The equalization with the STF indeed leads to considerably reduced spatial dependency in  $H_{HRTF,EQ}$  and aims at minimizing the required order for the SH transform. As the equalization dataset  $H_{EQ}$  can be described analytically, it can be determined at a freely chosen maximal order, typically  $N_{high} \geq 35$ . The SH coefficients for the equalized sparse HRTF set  $H_{HRTF,EQ}$  are obtained by applying the SH transform to the equalized HRTFs up to an appropriate low maximal order  $N_{low}$ , which corresponds to the maximal order that can be resolved by  $\Omega_s$ . Then an upsampled HRTF set  $\hat{H}_{HRTF,EQ}$  is calculated on a dense sampling grid  $\Omega_d = \{(\phi_1, \theta_1), \dots, (\phi_D, \theta_D)\}$ , with  $D \gg S$  by using the inverse SH transform. Finally, HRTFs are reconstructed by a subsequent de-equalization by means of spectral multiplication with a de-equalization dataset  $H_{DEQ}$

$$\hat{H}_{HRTF,DEQ}(\omega, \Omega_d) = \hat{H}_{HRTF,EQ}(\omega, \Omega_d) \cdot H_{DEQ}(\omega, \Omega_d). \quad (2)$$

For de-equalization, again the STF can be used. This last step recovers energies at higher spatial orders that were transformed to lower orders in the equalization.  $H_{HRTF} = \hat{H}_{HRTF,DEQ}$  holds if  $N_{low}$  and  $N_{high}$  are chosen appropriately. Energy which, after the equalization, still is apparent at high modal orders  $N > N_{low}$  is irreversibly mirrored to lower orders  $N \leq N_{low}$ . Thus we obtain  $H_{HRTF} \approx \hat{H}_{HRTF,DEQ}$ .

### 3. HRTF Datasets

We used HRTFs of a Neumann KU100 dummy head measured on a dense Lebedev grid with 2702 sampling points which can be used for SH processing at a sufficient order of  $N = 35$  for the evaluation [6]. The SH representation of the dataset served as the reference in our investigations. From this reference set we generated various sparse HRTF sets which were required as input data for the evaluation. First the sparse HRTF sets varied regarding the accessible spatial order  $N$ . These sets were obtained in the same way as described in [12] by spatially subsampling the reference set in the SH domain by means of the inverse SH transform. Furthermore, in order to create datasets incorporating positioning inaccuracies, we randomly varied the distance for each measured direction in a range of  $\pm \Delta r_{max}$  to the reference distance of  $R = 2m$ . To perform the distance shifts we used a method which is based on the SUPDEq method [2]. Instead of an incident plane wave (representing a sound source in the far-field), an STF for a spherical wave (point source) at the reference distance of  $R = 2m$  is used for the equalization and a spherical wave at  $R' = R + \Delta r$  for the de-equalization. By this, both the phase and the amplitude are appropriately adapted to the changed distance. To consider angular inaccuracies, we randomly modified the directions for which we determined the

HRTFs from the dense HRTF set equally-distributed within a solid angle of  $\Delta\phi_{max}$ .

It is worth noting that we chose randomly distributed deviations because they showed in informal pretests the highest impact on the spatial upsampling. Furthermore, such deviations are on the one hand typical, when the subject moves slightly between each of the sequentially measured directions or turns the head not exactly to the target direction. On the other hand measurement errors of a head-tracking device, used to determine the exact subject position and orientation during an HRTF measurement can be regarded as well being randomly distributed.

Accordingly we created datasets considering maximal distance deviations of  $\Delta r_{max} = 1cm, 2cm, 5cm$  and maximal angular deviations of  $\Delta\phi_{max} = 2^\circ, 5^\circ, 15^\circ$  by spatially downsampling of the reference set for 15 sparse sampling grids – Lebedev grids with 6, 14, 26, 38, 50, 74, 86, 110, 146, 170, 194, 230, 266, 302, and 350 sampling points – equaling (limited) orders of  $N = 1 - 15$ . For each of these conditions, we generated SH coefficients which we used for the further evaluation. Thus, both order-limited and de-equalized sets were always based on the respective sparse grid.

While the order-limited (OL) datasets were obtained with an SH interpolation without any pre- or postprocessing, we used the Matlab-based implementation of the SUPDEq method as described in [12] to obtain the de-equalized HRTF sets (DEQ). The radius for the rigid sphere model was calculated according to Algazi et al. [1] based on the dimensions of the dummy head, resulting in a radius of  $r = 9.19cm$ . Finally the HRTFs of the test grids used in the evaluations were obtained via the inverse SH transform of the order-limited dataset or the de-equalized dataset at the respective positions.

## 4. Evaluation

### 4.1. Spectrum

First we analyze the spectral deviations to the reference set as a function of  $N$  on a Lebedev grid with  $T = 2702$  sampling points as test sampling grid  $\Omega_t = \{(\phi_1, \theta_1), \dots, (\phi_T, \theta_T)\}$ . For this the frequency-dependent spectral differences per sampling point were calculated in dB as

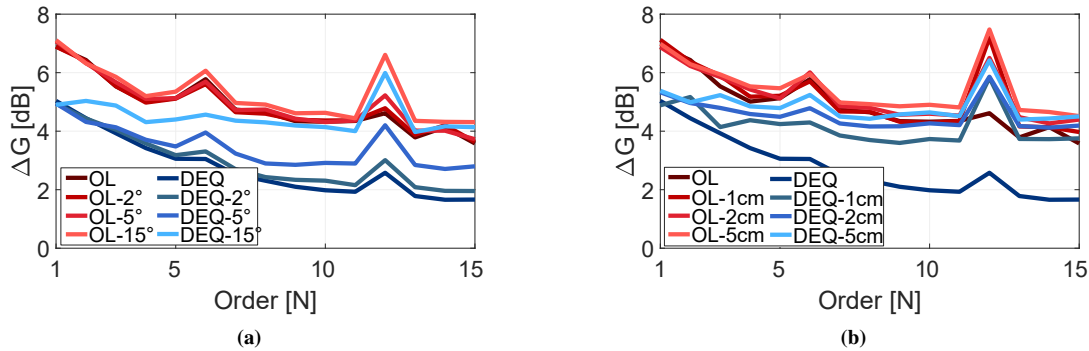
$$\Delta g(\omega, \Omega_t) = 20 \lg \left| \frac{H_{HRTF,REF}(\omega, \Omega_t)}{H_{HRTF,TEST}(\omega, \Omega_t)} \right|, \quad (3)$$

where  $H_{HRTF,REF}$  is the left ear HRTF extracted from the reference set and  $H_{HRTF,TEST}$  the one extracted from the order-limited or the de-equalized datasets at each sampling point  $\Omega_t$ . Then the absolute value of  $\Delta g(\omega, \Omega_t)$  was averaged across across all sampling points  $\Omega_t$  to obtain the frequency-dependent measure  $\Delta G_f(\omega)$  (in dB)

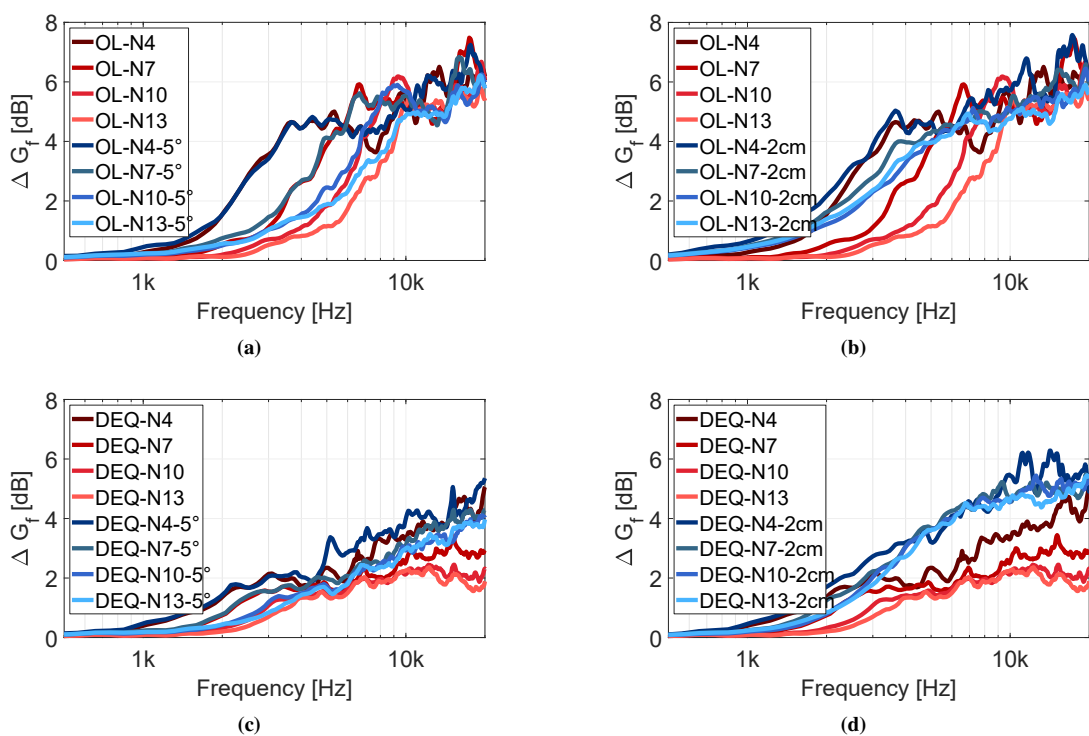
$$\Delta G_f(\omega) = \frac{1}{n_{\Omega_t}} \sum_{\Omega_t=1}^{n_{\Omega_t}} |\Delta g(\omega, \Omega_t)|, \quad (4)$$

and across  $\omega$  and  $\Omega_t$ , resulting in a single value  $\Delta G$  (in dB) describing the spectral difference

$$\Delta G = \frac{1}{n_{\Omega_t}} \frac{1}{n_{\omega}} \sum_{\Omega_t=1}^{n_{\Omega_t}} \sum_{\omega=1}^{n_{\omega}} |\Delta g(\omega, \Omega_t)|. \quad (5)$$



**Fig. 2:** Mean spectral differences  $\Delta G$  in dB (left ear) between reference HRTF set ( $N = 35$ ) and the datasets with angular and distances inaccuracies depending on the order  $N$ . Red: order-limited datasets (OL), Blue: de-equalized datasets (DEQ). (a) Influence of the angular inaccuracies  $\Delta\phi_{max}$ , (b) Impact of distance inaccuracies  $\Delta r_{max}$ . The color saturation corresponds to the size of the error.

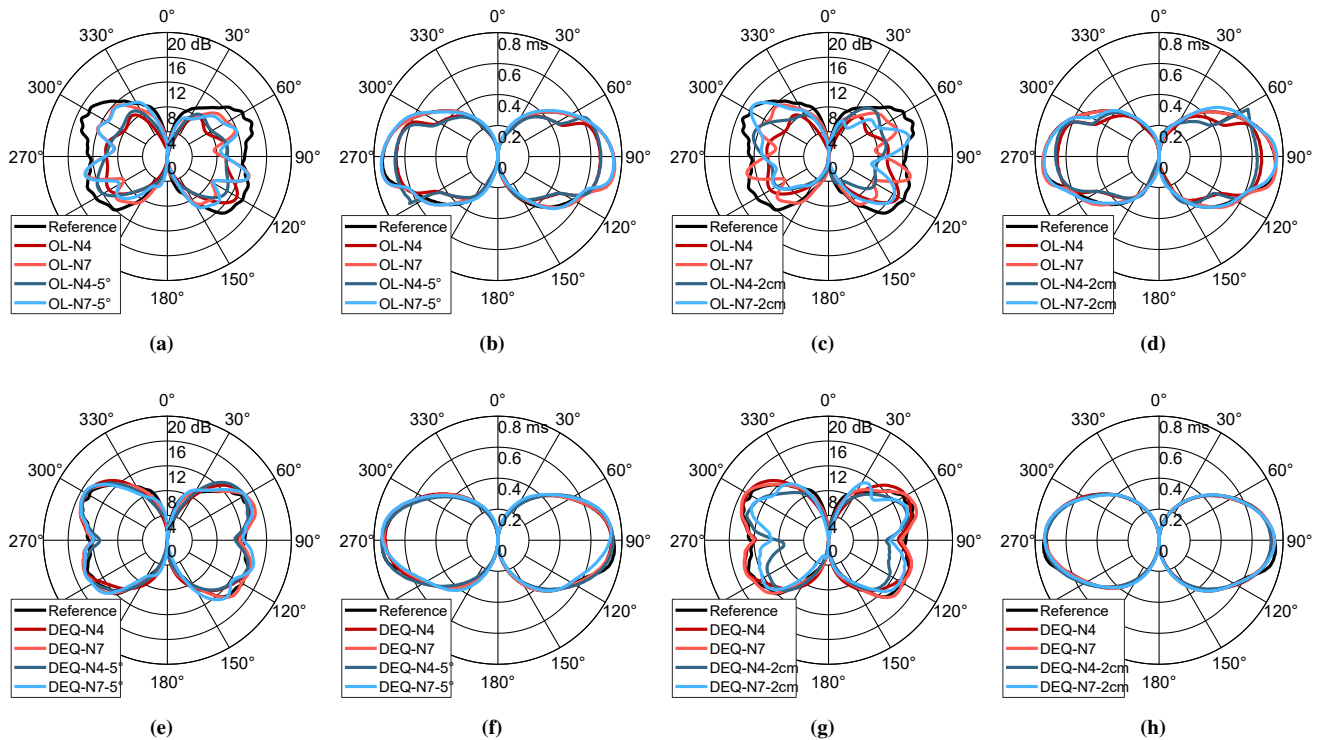


**Fig. 3:** Spectral differences  $\Delta G_f(\omega)$  in dB (left ear) of spatially upsampled datasets (color saturation corresponds to  $N$ ) to the reference HRTF set ( $N = 35$ ). Red: datasets without positioning inaccuracies, Blue: datasets with angular errors or distance inaccuracies. (a,b) Results for the order-limited sets (OL), (c,d) Results for the de-equalized sets (DEQ). (a,c) Influence of the angular alteration  $\Delta\phi_{max} = 5^\circ$ , (b,d) Influence of the distance shifts  $\Delta r_{max} = 2\text{cm}$ .

Fig. 2 shows  $\Delta G$  depending on the order  $N$  both for the order-limited datasets and the de-equalized datasets, that means without and with the SUPDEq processing. Data for angular deviations of  $2^\circ$ ,  $5^\circ$  and  $15^\circ$  as well as for distance deviations of  $1\text{cm}$ ,  $2\text{cm}$  and  $5\text{cm}$  is given. Generally, for the order-limited datasets, angular and distance deviations have only a minor influence on  $\Delta G$ . For de-equalized HRTF sets angular inaccuracies of up to  $\Delta\phi_{max} = 5^\circ$  result in a slight increase of  $\Delta G$ . Only for  $\Delta\phi_{max} = 15^\circ$  a strong influence already at low orders  $N$  can be observed (Fig. 2 a). On the contrary, distance inaccuracies strongly affect the de-equalized datasets (Fig. 2 b). Already for small  $\Delta r_{max}$ , the spectral differences increase, especially at higher spatial orders  $N$ . For example,

at  $N = 7$  and  $\Delta r_{max} = 1\text{cm}$  the increase is about 1.5 dB.

Fig. 3 shows the spectral differences over frequency. While the influence of angular deviations is minor for  $N = 4$  and  $N = 7$  it causes an increase for higher orders. However, the deviations are below 1 dB in all cases for frequencies up to 10 kHz. On the contrary, the influences of distance inaccuracies are much larger. Errors of  $\Delta r_{max} = 2\text{cm}$  strongly deteriorate the spectrum both for the de-equalized and the order-limited datasets. Furthermore, the inaccuracies nearly completely outweigh the benefit of the SUPDEq method, especially for higher orders. Thus, when performing HRTF measurements, distance inaccuracies between the sound source and the human head need to be avoided.



**Fig. 4:** ILDs and ITDs in the horizontal plane. Black: Reference HRTF set, Red: Datasets without positioning inaccuracies, Blue: Datasets with altered angles  $\Delta\phi_{max} = 5^\circ$  or distances  $\Delta r_{max} = 2\text{ cm}$ . In the upper line (a–d) the results for the order-limited datasets (OL) are shown, in the lower one (e–h) the ones for the de-equalized sets (DEQ). The angle represents the azimuth  $\phi$  of the sound source. The radius describes the magnitude of the level differences (in dB) or time differences (in ms). The left two rows (a,b,e,f) shows the results for the angular inaccuracies and the right rows (c,d,g,h) the ones for the distance inaccuracies.

## 4.2. Binaural cues

In Fig. 4 the interaural level differences (ILDs) and the interaural time differences (ITDs) are shown. For the order-limited sets (Fig. 4 (a–d)), differences of the ITDs and ILDs to the reference vary most depending on the spatial order  $N$ . The maximal deviations from the reference are up to 4 dB for the order-limited sets at  $N = 7$  and more than 8 dB at  $N = 4$ . For the lateral sound incidence variations in ITDs of up to 0.1 ms at  $N = 4$  occur. While the influence of the angular inaccuracies is minor, the distance error affects the ILDs, but however does not generally lead to a strong increase of the error. As shown in Fig. 4 (e–h) for the de-equalized datasets (DEQ), the ITDs are only slightly affected by the positioning inaccuracies. On the contrary, the considered distance inaccuracies of 2 cm strongly influence the ILDs and lead for lateral sound incidence to deviations of 4 dB at  $N = 4$  and of about 2 dB at  $N = 7$ .

## 4.3. Localization performance

To analyze the impact of the distance and angular inaccuracies on localization performance in the median sagittal plane, we used the model from Baumgartner et al. [3]. The model compares the spectral structure of a reference HRTF set to a set of test HRTFs and calculates a probabilistic estimate of the perceived sound source location. Based on this estimate, the polar RMS error is determined which describes the expected angular error between the actual and perceived source positions. Additionally, it determines the quadrant error rate

specifying the rate of front-back or up-down confusions. To estimate the localization performance in the horizontal plane, we used the model from May et al. [10] which weighs the frequency-dependent binaural cues (ILDs, ITDs) to estimate the azimuthal position of a sound source based on a trained Gaussian mixture model. A lateral error can be calculated by comparing the intended and the estimated source position. We used the Auditory Modeling Toolbox (AMT) [17] for these calculations. The procedure for determining the errors has been described in detail in [12] and is in the following briefly outlined. We used a test sampling grid  $\Omega_t$  with  $\phi = \{0^\circ, 180^\circ\}$  and  $-30^\circ \leq \theta \leq 90^\circ$  in steps of  $1^\circ$  to estimate median plane localization performance and assumed a median listener sensitivity of  $S = 0.76$  (in accordance with Baumgartner et al. [3]). For the horizontal plane localization performance, we applied a test sampling grid with  $\phi = \pm 90^\circ$  in steps of  $5^\circ$ . We calculated the absolute polar error difference (PE in degree)

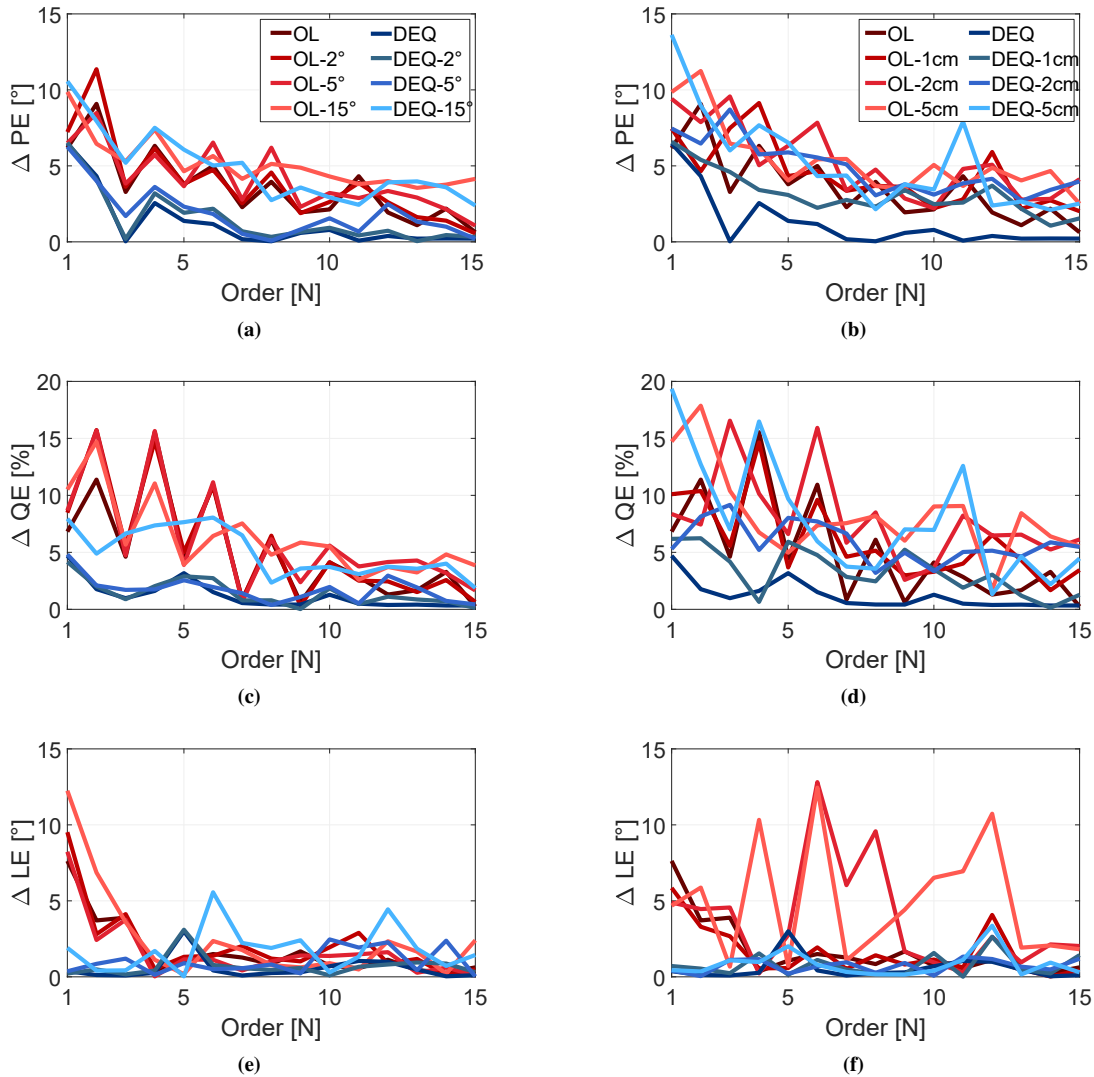
$$\Delta PE = |PE_{REF} - PE_{TEST}|, \quad (6)$$

the absolute quadrant error difference (QE in percent)

$$\Delta QE = |QE_{REF} - QE_{TEST}|, \quad (7)$$

as well as the absolute lateral error difference (LE in degree)

$$\Delta LE = \frac{1}{T} \sum_{t=1}^T |LE_{REF}(\Omega_t) - LE_{TEST}(\Omega_t)|, \quad (8)$$



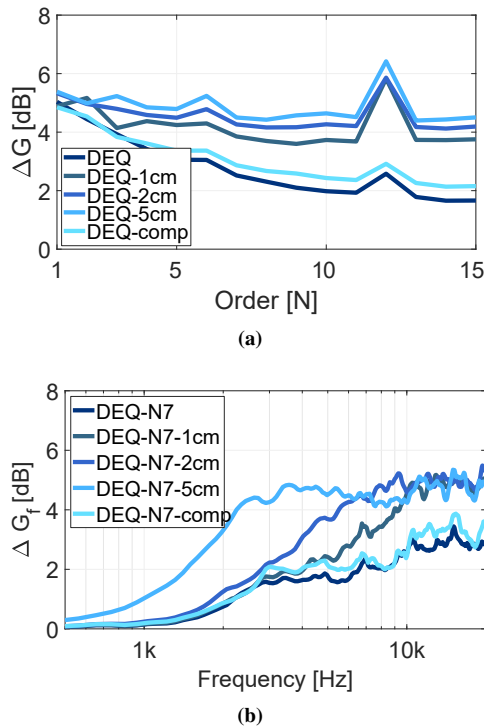
**Fig. 5:** Influence of distance and angular error on modeled localization performance. (a,b) Absolute polar error difference  $\Delta PE$ , (c,d) Quadrant error difference  $\Delta QE$ , (e,f) Lateral error difference  $\Delta LE$  over SH order  $N$ . Red: Results for the order-limited datasets (OL), Blue: De-equalized datasets (DEQ). Left row: (a,c,e) Influence of the angular inaccuracies  $\Delta\phi_{max}$ , Right row: (b,d,f) Impact of distance inaccuracies  $\Delta r_{max}$ . The color saturation corresponds to the size of the angular respectively the distance error.

for each order  $N$  with the subscripts *REF* for the reference set and *TEST* for the tested HRTF set.

As illustrated in Fig. 5 (a, c, e) angular inaccuracies only slightly affect localization performance of the order-limited datasets. For the de-equalized datasets the localization performance is strongly affected only for angular inaccuracies of  $\Delta\phi_{max} = 15^\circ$ . This is completely different for the distance inaccuracies which are shown in Fig. 5 (b, d, f). In the median sagittal plane already for  $\Delta r_{max} = 2\text{ cm}$ , both the polar error difference  $\Delta PE$  and the quadrant error difference  $\Delta QE$  are strongly increased. For some of the data, the localization errors are even stronger for the de-equalized datasets than for the order-limited datasets. Thus as already analyzed based on the spectral differences (see Sec. 4.1), distance inaccuracies have a severe impact on the spatial upsampling. However, the increase of the lateral error difference  $\Delta LE$  is quite small at least for the de-equalized datasets.

## 5. Distance error compensation

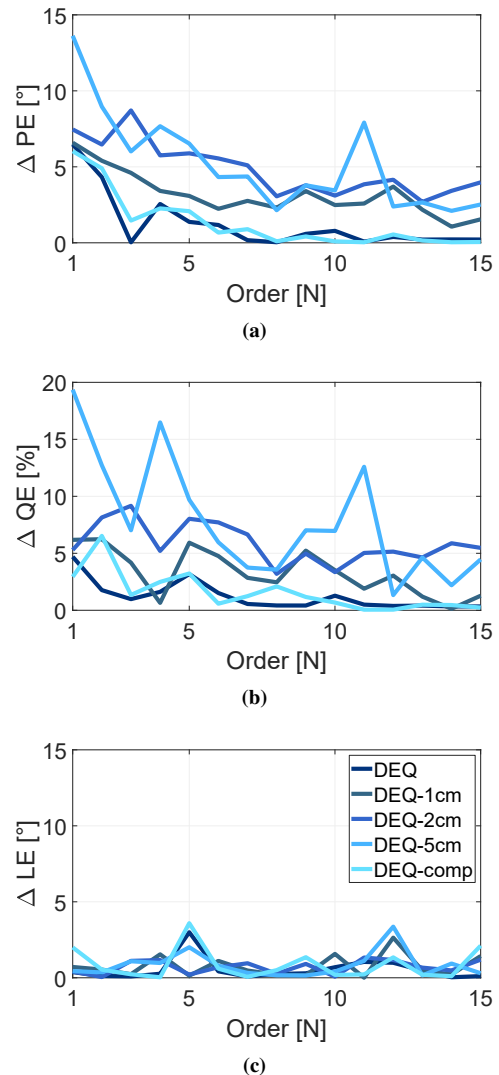
A major outcome of the present study is that already small (random) inaccuracies in the distance between sound source and listener strongly influence the spatial upsampling, especially when applying the SUPDEq method. As such positioning inaccuracies might not be systematic but somehow randomly distributed over the different measured directions they cannot be compensated using the recentering methods proposed e.g. by Richter et al. [16]. In the following we describe and examine the so-called DEC (Distance Error Compensation) method which reduces the influence of distance inaccuracies. The method is to some extent comparable to the approach from Ziegelwanger and Majdak [20] splitting up the HRTF in a direction-dependent part representing the influence of the sphere and direction-independent part. However, our implementation benefits from the directional equalization which is part of the SUPDEq method. As explained in Sec. 2, the spectral equalization (Eq. 1) removes direction-dependent



**Fig. 6:** Influence of the compensation on the upsampled HRTF sets on the mean spectral differences to the reference HRTF set ( $N = 35$ ). (a)  $\Delta G$  depending on the order  $N$ . (b)  $\Delta G_f(\omega)$  over frequency for  $N = 7$ . In addition to the compensated HRTF set (comp) the values for the measurement without positioning inaccuracies and for different distance inaccuracies  $\Delta r_{max}$  are given (color saturation).

spectral and temporal components from the measured HRTFs. By this head-related differences are removed in  $H_{HRTF, EQ}$  and ideally all peaks of the equalized head-related impulse responses (HRIR, the time-domain equivalent of an HRTF) are time-aligned. However, deviations due to the positioning inaccuracies remain after the equalization and thus the onset differences between the different equalized HRIRs directly relate to the distance errors. Thus, a simple onset-detection of the spatially equalized equalized HRIRs is applied to estimate the distance errors and to determine the required distance shift  $\Delta r$ . To apply the distance shift we use the directional equalization and subsequent de-equalization at different distances as already described in Sec. 3. For each measured HRTF a point source STF at the distance of  $R_{error} = R + \Delta r$  is used for the equalization and a point source STF at the reference distance of  $R = 2m$  for the de-equalization. After performing the DEC, the peaks of all equalized HRIRs are time-aligned. We implemented the DEC as a separate preprocessing step performed on the sparse HRTF set (see Fig. 1). A ten-times oversampling was applied for more precise onset detection and subsample accuracy. To be robust against noise we determined the onset based on -1 dB related to the maximal value of the equalized HRIRs.

The result of the compensation is given in Fig. 6 showing that the spectral differences are significantly reduced compared to the different distance inaccuracies  $\Delta r_{max}$  and are only slightly higher than for the dataset not comprising any positioning inaccuracies. The remaining differences are



**Fig. 7:** Influence of the compensation on the localization errors for the de-equalized datasets (DEQ). Absolute polar error difference  $\Delta PE$  (a), quadrant error difference  $\Delta QE$  (b), and lateral error difference  $\Delta LE$  (c) over SH order  $N$  for the compensated dataset (comp), the measurement without positioning inaccuracies and for varying distance inaccuracies  $\Delta r_{max}$  (color saturation).

mainly caused by the non-perfect directional equalization. Due to differences between the measured HRTF set and the equalization dataset some differences in the temporal structure remain, which are unintentionally included in the compensation. In Fig. 7 the impact of the distance error compensation on the localization performance is shown. The polar error difference  $\Delta PE$  (Fig. 7(a)), the quadrant error difference  $\Delta QE$  (Fig. 7(b)), and the lateral error difference  $\Delta LE$  (Fig. 7(c)) of the compensated datasets are as well very close to the sets without positioning inaccuracies. Thus an appropriate compensation of the distance inaccuracies as proposed in this paper can minimize localization errors of upsampled HRTF datasets.

## 6. Conclusion

In this paper we evaluated the influence of positioning inaccuracies of measured sparse HRTF sets on spatial upsampling.

For this we analyzed the impact of angular and distance errors on spectral cues, binaural cues and modeled localization performance for different spatial orders. The results can be summarized as follows. First, distance inaccuracies have much stronger impact than angular inaccuracies. Second, the effect of distance inaccuracies are much higher for the de-equalized sets than for the order-limited HRTF sets. Third, the influence of these inaccuracies becomes stronger with an increasing spatial resolution of the sparse HRTF set. Fourth, an appropriate distance error compensation (DEC), which applies a distance shift based on a time-alignment of the equalized HRIRs, can nearly completely eliminate the influence of distance inaccuracies. This way, distance inaccuracies in the HRTF sets have only minor impact on the spectral deviations and on the binaural cues. Examining modeled localization performance showed, that after the DEC nearly no influence of the investigated distance inaccuracies on localization performance remains.

Thus results show that the SUPDEq method can be used with simple and non-optimal measurement equipment incorporating angular and distance inaccuracies. Of course these findings need to be validated perceptually and based on measured individual datasets. To further validate the applicability of the method, a demonstration system needs to be set up allowing to measure spherical sparse HRTF sets.

The research presented in this paper has been funded by the German Federal Ministry of Education and Research. Support Code: BMBF 03FH014IX5-NarDasS. A Matlab-based implementation of the SUPDEq method is available on <https://github.com/AudioGroupCologne/SUPDEq>.

## 7. References

- [1] V. Algazi, C. Avendano, and R. O. Duda. Estimation of a Spherical-Head Model from Anthropometry. *J. Audio Eng. Soc.*, 49(6):472 – 479, 2001.
- [2] J. M. Arend and C. Pörschmann. Synthesis of Near-Field HRTFs by Directional Equalization of Far-Field Datasets. In *Proceedings of the 45th DAGA*, pages 1454–1457, 2019.
- [3] R. Baumgartner, P. Majdak, and B. Laback. Modeling sound-source localization in sagittal planes for human listeners. *J. Acous. Soc. Am.*, 136(2):791–802, 2014.
- [4] I. Ben Hagai, M. Pollow, M. Vorländer, and B. Rafaely. Acoustic centering of sources measured by surrounding spherical microphone arrays. *J. Acous. Soc. Am.*, 130(4):2003–2015, 2011.
- [5] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely. Spectral equalization in binaural signals represented by order-truncated spherical harmonics. *J. Acous. Soc. Am.*, 141(6):4087–4096, 2017.
- [6] B. Bernschütz. A Spherical Far Field HRIR / HRTF Compilation of the Neumann KU 100. In *Proceedings of the 39th DAGA*, pages 592–595, 2013.
- [7] B. Bernschütz, A. Vázquez Giner, C. Pörschmann, and J. M. Arend. Binaural reproduction of plane waves with reduced modal order. *Acta Acustica united with Acustica*, 100(5):972–983, 2014.
- [8] J. Blauert. *Spatial Hearing - The Psychophysics of Human Sound Localization*. MIT Press, Cambridge, MA, revised edition, 1996.
- [9] F. Brinkmann and S. Weinzierl. Comparison of head-related transfer functions pre-processing techniques for spherical harmonics decomposition. In *Proceedings of the AES Conference on Audio for Virtual and Augmented Reality*, pages 1–10, 2018.
- [10] T. May, S. Van De Par, and A. Kohlrausch. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(1):1–13, 2011.
- [11] C. Pörschmann and J. M. Arend. Obtaining Dense HRTF Sets from Sparse Measurements in Reverberant Environments. In *Proceedings of the AES Conference on Immersive and Interactive Audio*, 2019.
- [12] C. Pörschmann, J. M. Arend, and F. Brinkmann. Directional Equalization of Sparse Head-Related Transfer Function Sets for Spatial Upsampling. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 27(6):1060 – 1071, 2019.
- [13] C. Pörschmann, J. M. Arend, and F. Brinkmann. Spatial upsampling of individual sparse head-related transfer function sets by directional equalization. In *Proceedings of the 23rd International Congress on Acoustics*, 2019.
- [14] B. Rafaely. Analysis and Design of Spherical Microphone Arrays. *IEEE Trans. on Speech and Audio Proc.*, 13(1):135–143, 2005.
- [15] B. Rafaely. *Fundamentals of Spherical Array Processing*. Springer-Verlag, Berlin Heidelberg, 2015.
- [16] J. G. Richter, M. Pollow, F. Wefers, and J. Fels. Spherical harmonics based hrtf datasets: Implementation and evaluation for real-time auralization. *Acta Acustica united with Acustica*, 100(4):667–675, 2014.
- [17] P. Søndergaard and P. Majdak. The Auditory Modeling Toolbox. In J. Blauert, editor, *The Technology of Binaural Listening*, pages 33–56. Springer-Verlag, Berlin Heidelberg, 2013.
- [18] E. G. Williams. *Fourier Acoustics - Sound Radiation and Nearfield Acoustical Holography*. Academic Press, London, UK, 1999.
- [19] M. Zaunschirm, C. Schoerhuber, and R. Hoeldrich. Binaural rendering of Ambisonic signals by HRIR time alignment and a diffuseness constraint. *J. Acous. Soc. Am.*, 143(6):3616 – 3627, 2018.
- [20] H. Ziegelwanger and P. Majdak. Modeling the direction-continuous time-of-arrival in head-related transfer functions. *J. Acous. Soc. Am.*, 135(3):1278–1293, 2014.





## Full Reviewed Paper at ICSA 2019

Presented by VDT.

### Listening Tests with Individual versus Generic Head-Related Transfer Functions in Six-Degrees-of-Freedom Virtual Reality

Olli S. Rummukainen, Thomas Robotham, Axel Plinge, Frank Wefers,  
Jürgen Herre, and Emanuël A. P. Habets

*International Audio Laboratories Erlangen<sup>1</sup>, Germany, email: olli.rummukainen@iis.fraunhofer.de*

#### Abstract

Individual head-related transfer functions (HRTFs) improve localization accuracy and externalization in binaural audio reproduction compared to generic HRTFs. Listening tests are often conducted using generic HRTFs due to the difficulty of obtaining individual HRTFs for all participants. This study explores the ramifications of the choice of HRTFs for critical listening in a six-degrees-of-freedom audio-visual virtual environment, when participants are presented with an overall audio quality evaluation task. The study consists of two sessions using either individual or generic HRTFs. A small effect between the sessions is observed in a condition where elevation cues are impaired. Other conditions are rated similarly between individual and generic HRTFs.

#### 1. Introduction

Our ability to localize sounds in a 3-dimensional space relies on acoustic cues of interaural time difference (ITD), interaural level difference (ILD), interaural cross-correlation (ICC), and the spectral filtering caused by the physiology of the outer ears, head, and torso [14]. To render virtual audio binaurally over headphones, these cues are usually generated by convolving signals with head-related transfer functions (HRTFs). The HRTFs are individual; no two sets are alike.

Measuring or modeling individual HRTFs is a time consuming process requiring specialized hardware and facilities [4,9]. This process is currently unfeasible for the general public and most virtual experiences rely on non-individual, generic, HRTFs measured from a binaural head and torso simulator or averaged over a set of people. However, using generic HRTFs may result in blurry localization and front-back confusions, which in turn reduce the immersiveness of a virtual reality (VR) experience [25]. This study explores the consequences

of the choice between individual and generic HRTFs in a six-degrees-of-freedom (6-DoF) virtual environment.

In VR, in contrast to purely auditory research and applications, multiple modalities help us to construct a mental representation of our surroundings [7]. Visual information improves sound localization in real and virtual environments [1]. Freedom of movement in VR further improves our ability to extract and disambiguate sensory information. Sensory-motor coupling has been argued to form a basis for human cognition, where motor actions support sensory information processing [6]. This is clearly demonstrated by the reduction of front-back confusions when head movements are allowed [11, 24]. On the contrary, when we have 6-DoF full-body motion in a VR environment, the auditory localization resolution was found to be degraded when compared to stationary listening in a recent study [20]. This finding hints towards sound localization being less fine grained regardless of whether we are using individual or generic HRTFs during full-body motion.

A number of studies have looked into adaptation to altered

<sup>1</sup>A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS).

sound localization cues due to ear molds, hearing aids, or non-individual HRTFs via different training procedures. Active learning with feedback has been found to improve generic HRTF localization accuracy [13]. In VR, audio-visual cross-modal training improves source localization accuracy already after short training periods [3]. A many-to-one mapping mechanism in sound localization has been suggested, where the plasticity in the auditory cortex allows us to learn multiple HRTF sets that lead to the same percept [23]. For additional studies on auditory space adaptations, we refer the reader to [12].

A somewhat different approach is to select perceptually best-matching generic HRTFs from a large pool of HRTFs. Some participants have been shown to benefit from playing a VR-shooter game with their perceptual best-match HRTFs [17]. Related to adaptation, humans may adapt even to their perceptually worst-matched HRTFs through repeated short (12 min) training sessions [22]. However, the adaptation did not happen for everyone, suggesting that there may be limitations to adaptation in the worst match cases. No identifier was found to predict the individual ability to adapt to a new set of HRTFs. Anthropometry-based HRTF matching has been investigated in the context of VR, where no effect on questionnaire results was found between the best match HRTFs and generic HRTFs for a free exploration task in a VR scene [21].

Most studies on HRTF individualization have focused on localization accuracy only. Both timbral and spatial aspects of HRTF preference were considered in [2], where surprisingly a general preference of generic HRTFs over individual HRTFs was found. Potential causes were identified as higher quality of the built-in microphones of the generic binaural head and possible participant movement during HRTF measurement. These findings stress the importance to consider also non-localization-based HRTF quality attributes.

In this study the effect of individual versus generic HRTFs is investigated in the context of 6-DoF audio-visual virtual reality. In contrast to previous studies which have largely focused on localization accuracy with limited freedom of movement, the focus is on overall quality of audio rendering given self-movement cues and a corresponding visual environment. We examine the effect of the HRTF set with similar conditions and same participants in two separate sessions, namely, one with individual and one with generic HRTFs. The conditions include purposefully delayed tracking data and impaired localization cues in addition to a high quality convolution-based rendering and low quality non-spatial anchor conditions. We hypothesize  $\mathcal{H}_1$ : The individual HRTF session results in lower scores for some impairment conditions compared to the generic HRTF session,  $\mathcal{H}_2$ : The individual HRTF session results in less variance in the scores compared to the generic HRTF session, and  $\mathcal{H}_3$ : Individuals whose HRTFs are less like the generic HRTFs show more separation in scores between the sessions.

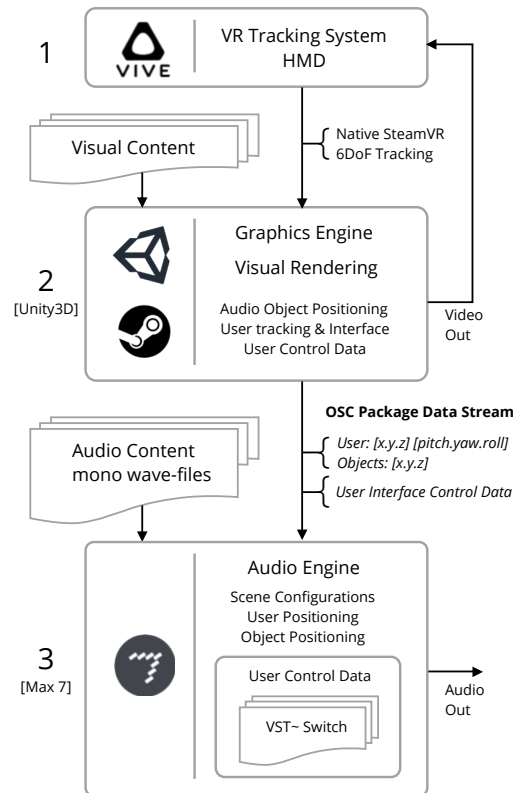


Fig. 1: Overview of the real-time evaluation platform.

## 2. Method

### 2.1. Virtual reality environment

A platform for real-time evaluation of binaural renderers in virtual reality was employed. The platform allows participants to switch between conditions, with no interruption to audio-visual sensory input, while exploring a 6-DoF virtual environment. The basic structure is presented in this section; for a thorough walk-through, please see [19]. The platform may be broken up into three components: 1) VR device, 2) Graphical rendering engine, and 3) Audio rendering engine, as depicted in Figure 1.

For Component 1, the HTC VIVE Pro<sup>1</sup> head-mounted display (HMD) is used for positional tracking, visual presentation, and control interface. The tracking accuracy and latency are found suitable for reproducible scientific research [15]. For Component 2, the Unity3D game engine is used for graphical rendering, along with hosting positional information for all audio objects and participant’s position and orientation using the SteamVR asset. All relevant positional and rotational data is then sent (via an Open Sound Control (OSC) data package at a 10 ms interval) to Component 3. In Component 3, a binaural renderer is hosted in Max 7 and is fed the positional and rotational information received from Unity3D. All audio content is loaded into Max on a scene by scene basis and triggered to play when the respective scene is loaded inside Unity3D.

<sup>1</sup><https://www.vive.com/eu/product/vive-pro/> (Accessed: 29.05.2019)



**Fig. 2:** Graphical user interface activated within the virtual environment and the interaction device.

As participants’ location is non-static, the test control interface is implemented inside the VR environment itself, allowing full freedom of movement while not being forced to return to a specific location to interact with the experiment interface. The interface is designed such that it can be instantiated anywhere in the VR scene. By pressing a button on a hand-held controller, a semi-transparent panel appears at eye level in the participants’ field of view. The panel is presented in Figure 2. Pressing the button again hides the panel, allowing the user to fully explore the environment. When instantiated, a virtual laser pointer may be used to operate sliders and buttons on the panel.

The participants have a possibility to teleport in the virtual scene, which means they are not bound by the tracked lab area. They have a 2 m × 3 m floor area for free walking, but this area may be re-positioned in the virtual world via the teleport function.

## 2.2. Participants

In total 10 people (2 female, 8 male) participated in the test. Their average age is 36.6 years (SD = 9.2). All the participants are doctoral students or employees at Fraunhofer IIS working in audio, and all of them have prior experience of VR systems. The participants may be considered as experts in audio quality evaluation, but none had experience on quality evaluation in VR context and they received no information about the conditions under test. Some of the participants had prior listening experience with the generic HRTFs used in this study. None reported any known hearing impairments.

## 2.3. Stimuli

**Scenes** There were four scenes with different characteristics used in the study. The scenes and the audio objects in the scenes are summarized in Table 1. The scenes were always evaluated in the same order as follows: *Restaurant* scene with three audio sources close to the horizontal plane, *Living room* with a fireplace audio object on the floor and a piano and wind chimes objects positioned at the same horizontal position but separated in elevation, *Outdoor* scene with bird sounds, tree cutting, and an airplane at different elevations, and *Fountain music* with a piano and a water fountain audio object.

**Tab. 1:** Visual scene and audio object descriptions.

Scene	Features
Restaurant	Three audio objects: Guitar, conversation, and bottle opening
Living room	Three audio objects: Piano, wind chimes, and fireplace
Outdoor	Four audio objects: Tree cutting, ducks, airplane, and birds
Fountain music	Two audio objects: Piano and a fountain

The audio objects in the three first scenes were presented visually as yellow spheres at the location of the audio event and there was no semantic congruency otherwise. In the fourth scene, *Fountain music*, the sound producing objects were visually modeled according to their real world counterparts as a piano and a fountain. All audio objects were initially within sight of the participant at the start of a specific scene. The participants could then teleport or walk closer to the objects to examine different aspects of the scene more carefully.

The audio samples were constructed of about one minute long segments that could be looped infinitely. They were recorded at 48 kbps with 24 bits. The acoustics of the virtual space was not modeled in the rendering stage. Some of the audio samples contained a small amount of reverberation from the recording location, but, due to the lack of virtual acoustics, the direct-to-reverberant ratio cue was not modeled. Thus, distance rendering relied only on the intensity cue realized by applying the inverse square law with a maximum level reached at 0.1 m from the sound object. Auditory near-field effects were not modeled. The audio was played back through Beyerdynamic DT770 Pro headphones and there was no individual headphone equalization applied. The headphones are diffuse-field equalized by the manufacturer.

**HRTFs** Individual HRTFs were measured in a semi-anechoic chamber at 1.2 m distance with a procedure described in [18]. The measurement setup is depicted in Figure 3, where the loudspeaker arc used for measurement signals is shown behind a participant who is standing on a rotating platform. The resolution was 5° in azimuth and 2.5° in elevation resulting in 4608 source positions with a filter length of 386 samples. The measurements were made at the entrance to the blocked ear canal. To remove any non-directional characteristics, the HRTFs were diffuse-field equalized for this experiment by calculating the average magnitude over all directions, inverting it, and creating a minimum-phase filter, which was then convolved with the HRTFs. Diffuse-field equalization through the recording stage to reproduction has been found to result in consistent playback for different listeners and to reduce the need for individual headphone equalization [10].

The generic binaural head (Neumann KU100) HRTFs were obtained from the spatial audio for domestic interactive en-





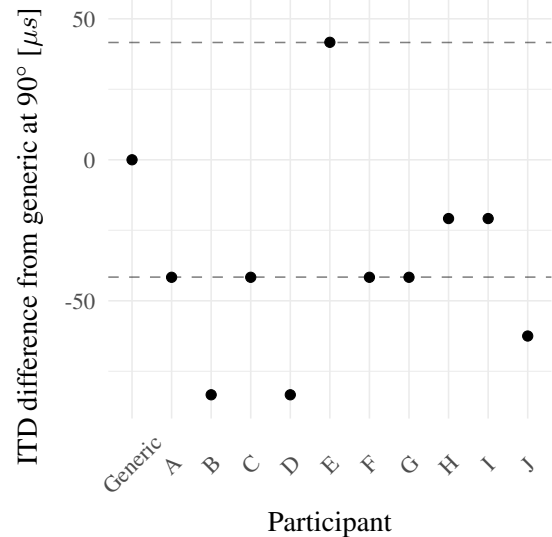
**Fig. 3:** HRTF measurement setup with a participant.

tainment (SADIE) database<sup>2</sup>. The measurements comprise 1550 source positions measured at 1.5 m distance and the resulting HRTFs are diffuse-field equalized. Filter length is 256 samples. Notably, the KU100 binaural head does not include a torso, removing the shoulder reflection effect from the generic HRTFs. The generic and individual HRTF sets were level aligned to 68 dB<sub>A</sub> at 1 m distance using pink noise and a binaural head.

To characterize the HRTFs objectively, the ITD was estimated from the HRTF sets for a source position directly to the side and at ear level. The estimation was based on finding the lag value corresponding to the maximum peak in the cross-correlation between the left and right ear head-related impulse responses. This value correlates with the individual head size. The resulting estimated differences between the generic HRTFs and the 10 participating individuals are displayed in Figure 4. Additionally, Figure 5 displays the diffuse-field equalized frequency responses of the individual HRTFs together with the generic HRTF for a source in the median plane at 40° elevation. The generic HRTF is observed to display one major notch between 7 kHz and 8 kHz, whereas the individual HRTFs have possibly multiple sharp notches in the range between 7 kHz and 14 kHz. Overall, the generic HRTF shows less detail in the frequency domain compared to the individual HRTFs, which results most likely from the shorter filter length (256 vs. 386 samples) and the simplified and torso-less geometry of the KU100 binaural head.

**Conditions** There were five audio rendering conditions in the test. The first, *Convolution*, convolved the nearest pair of HRTFs for each audio object’s direction of arrival with the

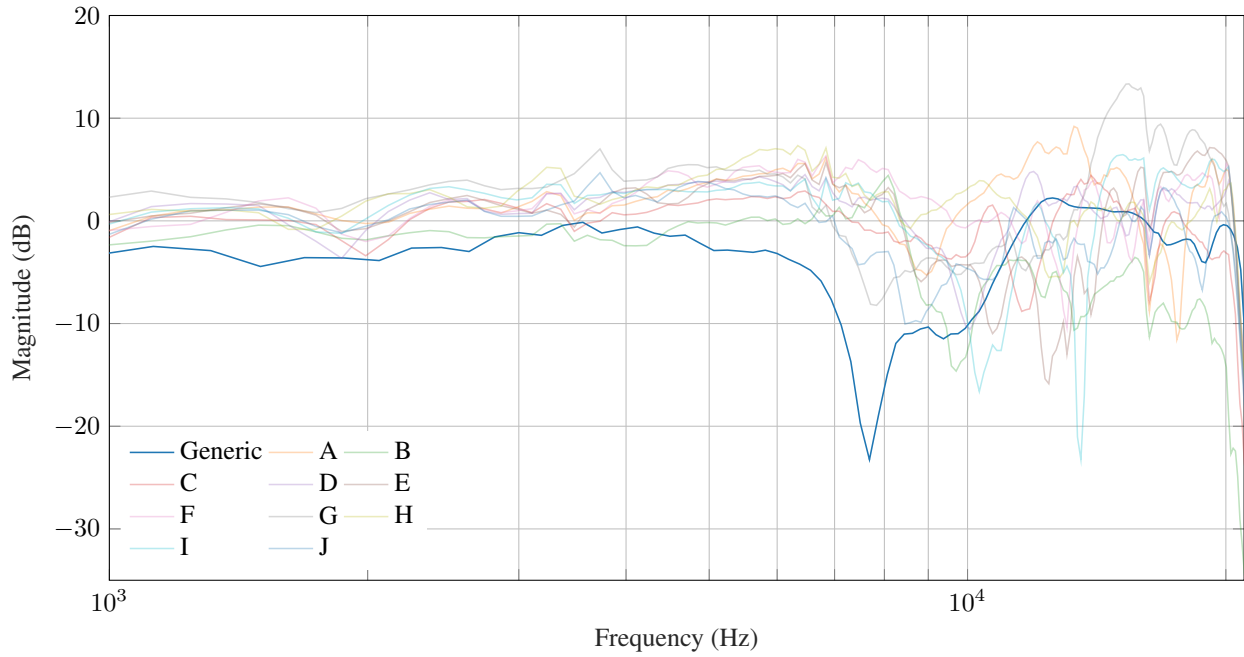
<sup>2</sup><https://www.york.ac.uk/sadie-project/index.html> (Accessed: 29.05.2019)



**Fig. 4:** ITD differences between the generic and the 10 individuals estimated from the HRTFs of the generic head (KU100) and the participants at azimuth = 90°, elevation = 0°. The dashed lines at 42 μs denote the boundaries of *low* and *high* similarity to the generic ITD groups, which are employed in the data analysis.

source signal. There was no interpolation between positions, i.e., the nearest HRTFs would always be selected. This condition functions as a basis for all the subsequent conditions, which were constructed by manipulating the tracking data or the scene setup. The *No elevation* condition modified the scene setup by placing all the audio objects’ elevation to the ear level based on the tracking data of each participant. Here, the visual scene remained unaltered, but the corresponding binaural rendering emanated from the ear level at the correct azimuthal direction. Similarly, the *Angle offset* condition biased the horizontal tracking data to shift all audio rendering by 15° counter-clockwise. In this condition the elevations were rendered correctly. These two conditions resulted in a constant audio-visual mismatch in location, that became increasingly evident the closer one is to the audio object. The *Delay 500 ms* condition added a delay of 500 ms to all tracking data for the audio rendering. Effectively, this condition resulted in a sound scene that reacted slowly to listener movements in position and rotation. Finally, the *Stereo mix* was not reactive to the listener movements, rather all audio objects were rendered in a static manner and without any HRTF processing. This condition resulted in a sound scene that is mostly localized within the head. The conditions are summarized in Table 2.

It was assumed that the degradation in elevation cues would most likely lead to differences between the sessions. The generic HRTFs are known to result in poor localization in elevation, which, in turn, could lead to a stronger audio-visual integration effect. Thus, there could be less reduction in overall quality scores because the visual target would capture the weakly localized auditory percept. The individual HRTFs would lead to stronger auditory localization and the audio-visual mismatch would be more easily perceived.



**Fig. 5:** Diffuse-field equalized frequency responses of the generic and individual head-related transfer functions measured at the left ear with azimuth = 0°, elevation = 40°.

**Tab. 2:** Audio rendering conditions under study.

Condition	Description
Convolution	HRTF-convolution-based renderer
No elevation	All audio objects are re-positioned to ear level
Angle offset	15° offset to the azimuth tracking data
Delay 500 ms	500 ms delay added to all tracking data
Stereo mix	Static stereo mix of all audio objects in a scene

## 2.4. Procedure

The participants were instructed to rate the overall quality of audio in the VR scene on a 100-point continuous scale in a multi stimulus test with hidden reference and anchor (MUSHRA) -like paradigm [8]. There were verbal labels marking the regions of the scale as bad, poor, fair, good, and excellent in 20-point intervals. It was made clear that there is no audio reference, but their self-motion and the visual presentations should be understood as creating the reference for expected auditory stimulation. Quality degradations were assumed to result from mismatch between the expectations and the perceived auditory stimulus. When unable to judge the overall quality, the participants were instructed to pay attention to spatial impression, i.e., how well the auditory stimulus is co-located with the visual sensation. The interface (Figure 2) allowed the participants to switch between the five conditions via buttons (A-E) as many times as required. There was no possibility to set loop points to inspect specific segments of the audio samples.

The participants were first familiarized with the VR system

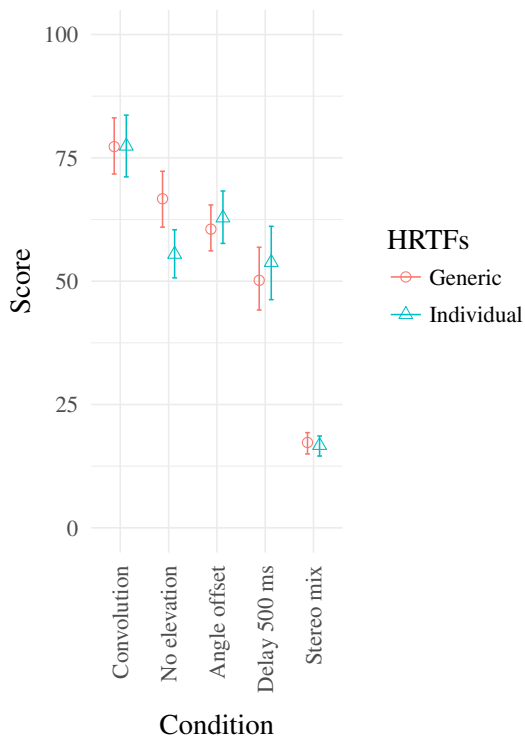
and interface in a special scene before beginning with the actual experiment. They were instructed on how to operate the controller to bring up and hide the control panel. They were also able to go through a dummy rating test without audio to get used to the interface buttons. The four experiment scenes were evaluated after the familiarization scene. There were two sessions conducted on separate days: the *individual* or *generic* HRTFs sessions. The conditions and scenes were identical in both sessions. The order of the session was counterbalanced between the participants so that half started with their own HRTFs and the other half with the generic HRTFs to remove the effect of learning. The average time to setup, familiarize, and complete evaluation of the four scenes in one session was 20 min.

## 3. Results

The study was structured around three independent variables in a within-participants design: *Scene*, *Condition*, and *Session*. The main effects on the dependent variable *Score* were analyzed by a three-way repeated measures analysis of variance (ANOVA). To check for ANOVA’s assumptions, Mauchly’s test for sphericity was performed on the data and indicated that the assumption of sphericity had not been violated for any independent variable. No main effect of the *Scene* ( $F_{(3,27)} = 0.266, p = 0.850, \eta_G^2 = 0.000$ ) nor any significant interactions with other variables were found. Thus, the data from all scenes were pooled together.

The results presented in the following stem from a two-way repeated measures ANOVA. The significant main effects and post-hoc analysis are presented in Table 3. Post-hoc comparisons were performed with the Tukey’s method. Effect sizes are reported as the generalized eta-squared values, where

$\eta_G^2 = 0.01$  is considered a small effect,  $\eta_G^2 = 0.06$  a medium effect, and  $\eta_G^2 = 0.14$  a large effect [5, 16].



**Fig. 6:** Mean scores for conditions and sessions. The whiskers denote the bootstrapped 95 % confidence intervals of the mean.

Figure 6 displays the main results in a graphical form. Most importantly for our hypotheses, a significant interaction of *Session* and *Condition* is observed: in the *Generic* HRTF session the *No elevation* condition was scored significantly higher than in the *Individual* condition. Otherwise, both sessions result in similar scoring of the conditions. The unimpaired *Convolution* condition is always rated the highest and the *Stereo mix* the lowest. The *Delay 500 ms* condition receives the second lowest scores in both sessions followed slightly higher by the *Angle offset* and *No elevation*.

Further analysis on the variance between the sessions in each of the conditions was conducted by the Levene’s test of homogeneity of variance. The data was split according to the condition and a separate analysis on the effect of session was done on each sub-group. The Levene’s test revealed the homogeneity of variance assumption was met in every sub-group and thus no significant difference in the variances between sessions could be concluded based on our data. The analysis is in agreement with visual inspection of the distributions in Figure 6, where the 95 % confidence intervals are approximately similar for the two session in all conditions.

To test the hypothesis that the similarity of HRTFs to the generic ones has an effect, the participants were evenly divided into two groups based on their absolute ITD difference from the generic ITD. The split was done with ITD difference  $\leq 42 \mu s$  defined as the *high similarity* and ITD difference  $> 42 \mu s$  as the *low similarity* groups. Participants at the border were randomly assigned to either group to obtain even

groups. By visually inspecting the scores grouped by the ITD difference, no difference between the groups could be observed. The ITD difference was further added as a factor to a linear model to explain variation in the dependent *Score* variable, but no effect was found.

## 4. Discussion

The *individual* HRTF session was found to be scored lower in one impairment condition, *No elevation*, which lends support to our first hypothesis that the conditions would receive differing scoring between HRTF sessions. No differences were observed in any other condition between the sessions. The other impairment types were not as critical for accurate HRTFs, since they involved errors in horizontal localization, which is robust against spectral differences in the HRTFs, or delayed tracking data, which affects all localization.

The *No elevation* had a rather large impairment in the elevation cues especially in the *Outdoor* scene, where there were multiple elevated audio objects such as tree clipping, birds, and an airplane. In other scenes the participants could have moved close to an audio object and crouched below it to inspect the elevation rendering. This, however, depended on the participants since they received no special instruction what to do in the virtual environment. Similarly, it was easy to miss the  $15^\circ$  angular offset, thanks to the audio-visual integration.

In a 6 DoF VR, the participants may ultimately dictate the audio content by their movements. While the fundamental audio content within a virtual scene is the same, participants’ varying position and orientation means that audio in one participant’s experience will be different from another. This raises the question whether there should be a participant training program for VR audio quality evaluations and what would that program entail. On the one hand, there is value in evaluations done with naive participants, as they represent the average end-user of a VR service, but on the other hand, they may miss some obvious shortcomings in rendering that become evident by chance to someone else.

Our second hypothesis about the reduced variance in response scores for the *individual* HRTF session is not supported by the data. Inspecting the score distributions visually in Figure 6 the bootstrapped confidence intervals appear similar for both sessions. Furthermore, Levene’s test for homogeneity of variance did not find differences in the between sessions variances for any condition. Similarly, our third hypothesis concerning the differing scores based on the likeness of individual HRTFs to the generic HRTFs is rejected by our data. For these hypothesis our sample size (N=10) is probably too limited, since there is not enough variation in the HRTFs. The HRTF similarity comparison was based on ITD differences only, which may not be a descriptive or meaningful enough a metric.

In summary, our findings lend only weak support for the need of individual HRTFs for critical listening in 6-DoF VR. Only one effect between individual and generic HRTFs was observed with a small effect size ( $\eta_G^2 = 0.03$ ), signaling

**Tab. 3:** Main effects, interactions, generalized eta-squared ( $\eta_G^2$ ) effect sizes, and post-hoc comparisons with  $p < 0.05$ .

Effect	F-value	p-value	Effect size	Post-hoc ( $p < 0.05$ )
Condition	$F_{(4,36)} = 34.17$	$p < 0.001$	$\eta_G^2 = 0.63$	
Session	$F_{(1,9)} = 0.61$	$p = 0.46$	$\eta_G^2 = 0.00$	
Session $\times$ Condition	$F_{(4,36)} = 3.30$	$p = 0.02$	$\eta_G^2 = 0.03$	
Session No elevation	$F_{(1,9)} = 6.58$	$p = 0.03$		Individual < Generic
Condition Generic	$F_{(4,36)} = 34.22$	$p < 0.001$		Stereo mix < Delay < Angle offset < Convolution; Stereo mix < Delay < No elevation < Convolution
Condition Individual	$F_{(4,36)} = 25.82$	$p < 0.001$		Stereo mix < Delay < Angle offset < Convolution; Stereo mix < No elevation < Convolution

that errors in elevation rendering may go unnoticed with generic HRTFs. Further studies with larger sample sizes are needed to be able to draw a clearer image of the question. Furthermore, here the HRTF sets had a large difference in the number of source locations with 1550 (generic) versus 4608 (individual), which could be assumed to result in larger perceptual effects. In 6-DoF VR the visuals and self-movement potentially largely mask the reduced spatial resolution of the generic HRTFs through audio-visual-proprioceptive integration. However, our results do not mean that generic HRTFs will result in higher quality in 6-DoF VR compared to individual HRTFs. The individual and generic HRTFs were not directly compared against each other in this study, and most likely different results would emerge from such a comparison.

Finally, our observations point towards a need for a training session to inform participants of the different nature of impairments in 6-DoF audio. In informal discussions many participants commented having learned what to listen for only after the first session or during the second session. Taking previous literature on auditory adaptation into account, a training session in 6-DoF may also serve as a familiarization phase to the generic HRTFs further reducing the effect of individual versus generic HRTFs.

## 5. Conclusions

In this study the effect of individual versus generic HRTFs for critical listening was investigated in 6-DoF audio-visual virtual reality. The use of individual HRTFs was found to enhance the participants' perception of errors in elevation. However, the effect size was small and most of the conditions showed no difference between the HRTF sets. Future research directions are envisioned to include participant training in six-degrees-of-freedom VR audio evaluation and adaptation to generic HRTFs in 6-DoF VR.

## 6. References

- [1] AHRENS, A., LUND, K. D., MARSCHALL, M., AND DAU, T. Sound source localization with varying amount of visual information in virtual reality. *PLoS One* 14, 3 (2019), 1–19.
- [2] ARMSTRONG, C., THRESH, L., MURPHY, D., AND KEARNEY, G. A perceptual evaluation of individual and non-individual HRTFs : a case study of the SADIE II database. *Applied Sciences* 8, 2029 (2018), 1–21.
- [3] BERGER, C. C., GONZALEZ-FRANCO, M., TAJADURA-JIMÉNEZ, A., FLORENCIO, D., AND ZHANG, Z. Generic HRTFs may be good enough in virtual reality. Improving source localization through cross-modal plasticity. *Frontiers in Neuroscience* 12, February (2018).
- [4] CARPENTIER, T., BAHU, H., NOISTERNIG, M., AND WARUSFEL, O. Measurement of a head-related transfer function database with high spatial resolution. In *Forum Acusticum* (Krakow, Poland, 2014), pp. 1–6.
- [5] COHEN, J. *Statistical power analysis for the behavioral sciences*, 2nd ed. Routledge, New York (NY), USA, 1988.
- [6] ENGEL, A. K., MAYE, A., KURTHEN, M., AND KÖNIG, P. Where's the action? The pragmatic turn in cognitive science. *Trends in Cognitive Sciences* 17, 5 (2013), 202–209.
- [7] ERNST, M. O., AND BÜLTHOFF, H. H. Merging the senses into a robust percept. *Trends in Cognitive Sciences* 8, 4 (2004), 162–9.
- [8] INTERNATIONAL TELECOMMUNICATION UNION. Recommendation ITU-R BS.1534-3 Method for the subjective assessment of intermediate quality level of audio systems, 2015.
- [9] KATZ, B. F. G. Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation. *The Journal of the Acoustical Society of America* 110, 5 (2001), 2440–2448.
- [10] LARCHER, V., JOT, J.-M., AND VANDERNOOT, G. Equalization methods in binaural technology. In *Audio Engineering Society 105th Convention* (San Francisco (CA), USA, 1998), pp. 1–28.
- [11] MCANALLY, K. I., AND MARTIN, R. L. Sound localization with head movement: Implications for 3-d audio displays. *Frontiers in Neuroscience* 8 (2014), 1–6.



- [12] MENDONÇA, C. A review on auditory space adaptations to altered head-related cues. *Frontiers in Neuroscience* 8 (2014), 1–14.
- [13] MENDONÇA, C., CAMPOS, G., DIAS, P., VIEIRA, J., FERREIRA, J. P., AND SANTOS, J. A. On the improvement of localization accuracy with non-individualized HRTF-based sounds. *Journal of the Audio Engineering Society* 60, 10 (2012), 821–830.
- [14] MØLLER, H., SØRENSEN, M. F., HAMMERSHØI, D., AND JENSEN, C. B. Head related transfer functions of human subjects. *Journal of the Audio Engineering Society* 43, 5 (1995), 300–321.
- [15] NIEHORSTER, D. C., LI, L., AND LAPPE, M. The accuracy and precision of position and orientation tracking in the HTC Vive virtual reality system for scientific research. *i-Perception* 8, 3 (2017), 1–23.
- [16] OLEJNIK, S., AND ALGINA, J. Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological Methods* 8, 4 (2003), 434–447.
- [17] POIRIER-QUINOT, D., AND KATZ, B. F. G. Impact of HRTF individualization on player performance in a VR shooter game II. In *Audio Engineering Society Conference on Audio for Virtual and Augmented Reality* (Redmond (WA), USA, 2018), pp. 1–8.
- [18] RICHTER, J.-G., BEHLER, G., AND FELS, J. Evaluation of a fast HRTF measurement system. In *Audio Engineering Society 140th Convention* (Paris, France, 2016), pp. 1–7.
- [19] ROBOTHAM, T., RUMMUKAINEN, O., AND HABETS, E. A. P. Evaluation of binaural renderers in virtual reality environments: platform and examples. In *Audio Engineering Society 145th Convention* (New York (NY), USA, 2018), pp. 1–5.
- [20] RUMMUKAINEN, O. S., SCHLECHT, S. J., AND HABETS, E. A. P. Self-translation induced minimum audible angle. *The Journal of the Acoustical Society of America* 144, 4 (2018), EL340–EL345.
- [21] SIKSTRÖM, E., GERONAZZO, M., KLEIMOLA, J., AVANZINI, F., DE GÖTZEN, A., AND SERAFIN, S. Virtual reality exploration with different head-related transfer functions. In *15th Sound and Music Computing Conference* (Limassol, Cyprus, 2018), pp. 85–92.
- [22] STITT, P., PICINALI, L., AND KATZ, B. F. Auditory accommodation to poorly matched non-individual spectral localization cues through active learning. *Scientific Reports* 9, 1 (2019), 1–14.
- [23] TRAPEAU, R., AUBRAIS, V., AND SCHÖNWIESNER, M. Fast and persistent adaptation to new spectral cues for sound localization suggests a many-to-one mapping mechanism. *The Journal of the Acoustical Society of America* 140, 2 (2016), 879–890.
- [24] WALLACH, H. The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology* 27, 4 (1940), 339–368.
- [25] WENZEL, E. M., ARRUDA, M., KISTLER, D. J., AND WIGHTMAN, F. L. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America* 94, 1 (1993), 111–123.



## Full Reviewed Paper at ICSA 2019

Presented \* by VDT.

### Enhanced Immersion for Binaural Audio Reproduction of Ambisonics in Six-Degrees-of-Freedom: The Effect of Added Distance Information

Axel Plinge, Sebastian J. Schlecht, Olli S. Rummukainen, and Emanuël A. P. Habets  
*International Audio Laboratories Erlangen<sup>†</sup>, Germany*

#### Abstract

The immersion of the user is of key interest in the reproduction of acoustic scenes in virtual reality. It is enhanced when movement is possible in six degrees-of-freedom, i.e., three rotational plus three translational degrees. Further enhancement of immersion can be achieved when the user is not only able to move between distant sound sources, but can also move towards and behind close sources. In this paper, we employ a reproduction method for Ambisonics recordings from a single position that uses meta information on the distance of the sound sources in the recorded acoustic scene. A subjective study investigates the benefit of said distance information. Different spatial audio reproduction methods are compared with a multi-stimulus test. Two synthetic scenes are contrasted, one with close sources the user can walk around, and one with far away sources that can not be reached. We found that for close or distant sources, loudness changing with the distance enhances the experience. In case of close sources, the use of correct distance information was found to be important.

#### 1. Introduction

Immersion is one of the key goals of virtual reality (VR) [1]. Along with covering the field of view with live graphics, realistic spatial sound is an important component. It was found that tracked rendering improves localization [2, 3], meaning it is also important for realistic and immersive reproduction [4]. Thus tracked rendering is an active research topic in the VR community: It is important to know how accurate the reproduction has to be and how to get to that accuracy.

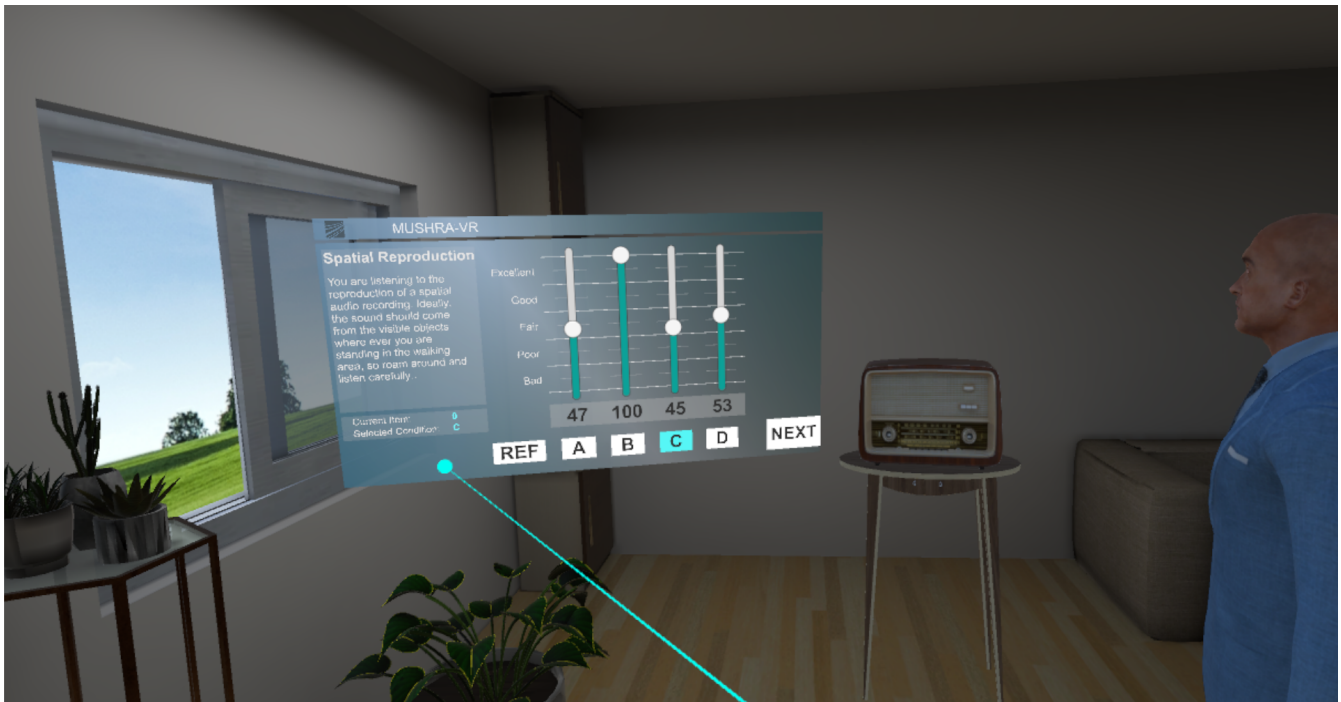
It was shown in our recent paper [5] that by using distance information in addition to spatial recording from one location allows six degrees-of-freedom (6DoF) reproduction. The directions of arrival are estimated from the recording, so with known distance the sound sources can be correctly positioned. The question investigated here is in which scenarios the added distance information leads to a better immersion. For spatial audio reproduction of recorded scenes, the fundamental pipeline is as follows: The recording is done from a

single or multiple spatially distributed positions with one or more microphone arrays; During reproduction, the listener's relative position is tracked and used to change the sound synthesis; The synthesis itself is done using loudspeakers or, in most cases, headphones. These three basic steps of the pipeline are briefly reviewed in the following sections.

##### 1.1. Recording

Regardless of the recording apparatus, spatial sound is often encoded in the Ambisonics format. This format is a compact and well defined representation of the sound field [6]. It uses the spherical harmonic domain representation [7], which captures the directional information of a sound scene with respect to a specific point in space. The distance of the points of origin of the sounds is not described. It is theoretically possible to derive distance information using the knowledge of the exact microphone positioning. However it might be hardly practical to derive the location of sound sources from the Ambisonics signal alone. The simplest way of interpreting

<sup>†</sup> Audiolabs is a joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS). \*



**Fig. 1:** VR scene 1, living room, with the MUSHRA panel. The sliders for rating the four conditions are in cyan, below them are the numerical value and a button each to select the condition. The active condition and a short descriptive text are shown on the left. The sound objects are the man on the left and the radio, which are inside the walking area of the listener. Bird sounds can be heard from the window.

the sound field for reproduction is to place the whole sound scene in the ‘far-field’, i.e. outside the reach of the listener.

The size of the recording apparatus has direct implications on the reproduction possibilities. Here we distinguish two cases: First, sound can be recorded either in a single location with compact array; Second, in multiple locations with distributed microphones or microphone arrays [8]. The first is typically the case for recordings done with consumer devices such as cameras and smartphone add-ons. Such a spatial recording from a single location is often used to reproduce a general spatial ambiance. Examples for the second case are dedicated recording sessions in sound or film studios, where multiple synchronized recording devices are employed. Such a recording allows for reproduction of a complex sound scene. It is possible to reproduce sound sources enclosed in the recording area by interpolation between the microphone arrays.

### 1.2. Spatial Processing

Next, we investigate some common processing methods for the reproduction of the spatial characteristics to a tracked listener. When the listener is moving, both the angle of arrival of the sound and its relative loudness change. Ideally, this is to be reproduced perfectly with perfect knowledge of the exact geometric location of each sound source and the acoustic absorption and reverberation effects of the environment.

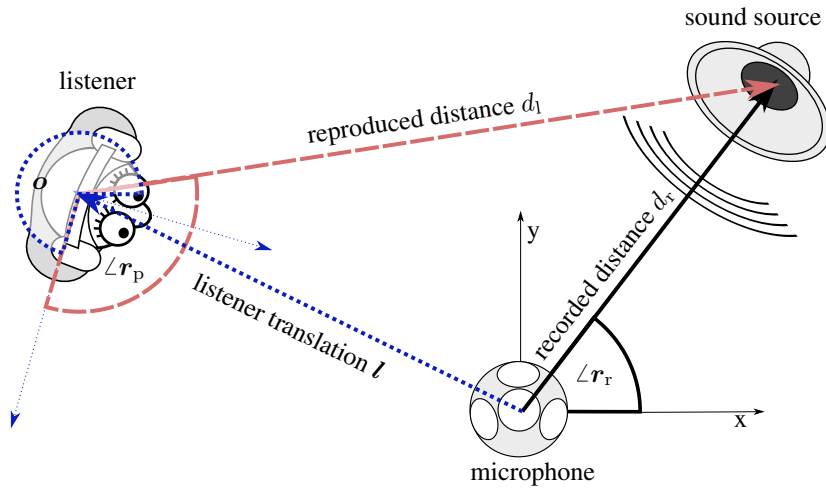
When reproducing a recording of a single compact microphone array, all sources are often placed in the far-field. Here, the relative distance of the listener in reproduction to the recorded sources is of limited consequence and the reproduction is often done in three degrees-of-freedom (3DoF) only. This means that only head rotation is applied. This rotation

can be computed directly inside the spherical harmonic domain [9].

It is, however, technically possible to extend the freedom of movement even if the recording was done in only a single location. This is facilitated by using parametric sound processing [10]. Examples of such manipulations are the so called ‘acoustic zoom’ techniques that allow the user to close in on a far-field sources in one direction [11, 12].

The novel method first presented in [5] brings the concept of extending the reproduction possibilities even one step further. It is also based on a single location of recording and parametric encoding of the recorded sound is employed to facilitate the necessary manipulations. By adding distance information for the sound source in each direction, it is possible to virtually place the sound sources in the room, allowing for full 6DoF. It thus is possible to walk around the sources in reproduction, cf. **Fig. 1**. This is a strong effect for immersion enhancement that will be investigated more thoroughly in this paper.

When multiple microphone arrays or Ambisonics microphones are used in a spatially distributed setup, a complex sound scene can be reproduced more easily. In order to enable the listener to walk around sound sources in the scene, the relative positioning of the recording devices has to be mapped to the reproduction scene. Then, the relative distance of the sound sources can be incorporated and the reproduction can be performed in full 6DoF. Different methods for interpolation of their signals have been proposed [13–16]. It is also possible to apply source separation techniques for isolating sound sources in the far-field, especially if the



**Fig. 2:** The 6DoF reproduction of spatial audio. A sound source is recorded by a microphone with the direction of arrival (DOA)  $r_r$  in the distance  $d_r$  relative to the microphones position and orientation (black line and arc). It has to be reproduced relative to the moving listener with the DOA  $r_l$  and distance  $d_l$  (red dashed). This has to consider the listeners translation  $l$  and rotation  $o$  (blue dotted).

recording is done with many microphones and then encoded in higher-order Ambisonics (HOA) [17].

### 1.3. Reproduction

Finally, we describe the actual reproduction. This can be implemented with loudspeakers or headphones.

Sound produced by loudspeakers is usually targeted at the so called ‘sweet spot’, a small area where the directions and loudness of all speakers are matched. The listener has to stay in this area in order to experience the intended spatial impression, with the head oriented in the intended direction. Some techniques allow for widening this area [18] or adjusting to head rotations [19]. The spatial reproduction ability is often limited by the loudspeakers’ arrangement, unless a large number of them is used.

The use of headphones is common in VR applications. The user is often wearing a head-mounted display (HMD), thus adding headphones is quite practical. When using headphones, head-related transfer functions (HRTFs) are applied in order to spatialize the sound [20, 21]. HRTFs allow to reproduce the effect of the shape of torso, head, and outer ears on the sound depending on its direction of origin. Often virtual loudspeakers are placed around the user. These virtual loudspeaker signals are then binauralized [22]. It is possible to binauralize Ambisonics by applying HRTFs directly in the spherical harmonic domain, however the limited Ambisonics order can lead to unwanted filtering effects in practice [23].

### 1.4. Research Question

The question this paper looks to answer is how important the distance information of sound sources is in VR reproduction from a recording at a single location. This will be done by a subjective listening experiment. Recently multi-stimulus testing has been employed in listening tests for VR. In several studies the multiple stimuli with hidden reference and anchor (MUSHRA) paradigm, common in general non-interactive audio testing, was applied [16,24]. The test subject is immersed in a VR scene, typically rendered as computer generated imaging (CGI). Different audio renderings can be

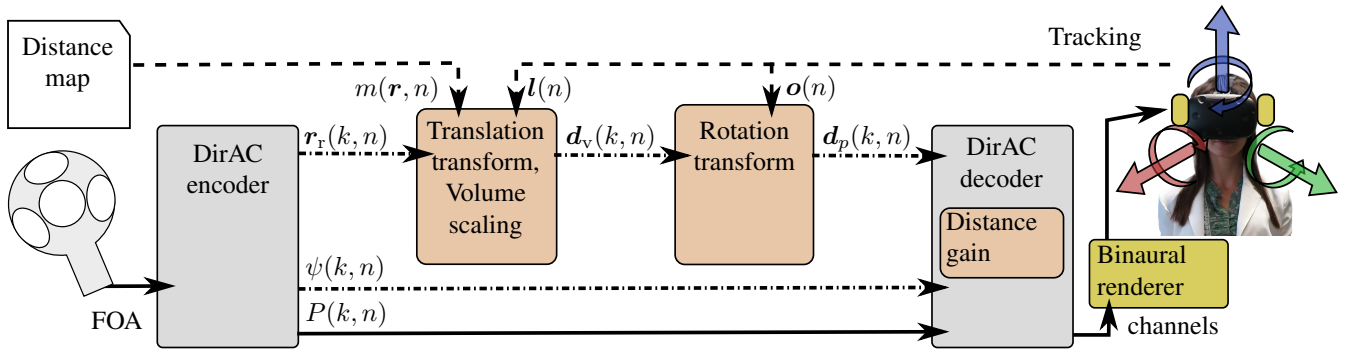
switched from inside the scene and rated on a scoreboard. In our experiment, such a MUSHRA test is performed in a virtual indoor scene as in our previous paper [5]. In this scene the listener can walk around the sources. This is contrasted with a second virtual outdoor scene, where the distance information is of less importance as the sources are placed outside the walking area. This way it is investigated when the use of distance information actually enhances the experience in a meaningful way. By keeping the processing pipeline the same for all conditions, the timbre is as close as possible. The only difference in conditions is the application of the tracking data to the sound processing. Angular and distance effects are separated to allow judging their individual importance.

The rest of this paper is organized as follows: First, the method introduced in [5] is described in Section 2. The individual implementation of the sound rendering and the scenes used are explained in Section 3. Section 4 gives the listening test results followed by a short conclusion in the final Section 5.

## 2. Method

In this paper, a binaural signal is produced at the listener’s position given the signal at a single recording position and information about the distances of sound sources from that recording position. Given a scene of limited size, the physical sources are assumed to be separable by their angle towards the recording position.

The position of the recording microphone is used as the origin of the reference coordinate system. The listener is tracked in 6DoF, cf. Fig. 2. At a given time, the listener is at a position  $l \in \mathbb{R}^3$  relative to the microphone and has a rotation  $o \in \mathbb{R}^3$  relative to the microphones’ coordinates system. We deliberately choose the recording position as the origin of our coordinate system to simplify the notation. The sound is reproduced with a different distance  $d_l$ , leading to a changed signal level, and a different DOA  $r_p$  that is the result of both translation and subsequent rotation.



**Fig. 3:** Proposed method of 6DoF reproduction. The recorded first-order Ambisonics (FOA) signal in B-format is processed by a directional audio coding (DirAC) encoder that computes direction and diffuseness values for each time-frequency bin of the complex spectrum. The direction vector is then transformed by the listener’s tracked position and according to the distance information given in a distance map. The resulting direction vector is then rotated according to the head rotation. Finally, signals for 8+4+4 virtual loudspeaker channels are synthesized in the DirAC decoder. These are then binauralized.

The reproduction pipeline introduced in [5] is sketched in **Fig. 3**. An Ambisonics recording and a distance map are used as input signal for rendering. In the parametric DirAC representation [25], geometric transformations are applied. Then a channel signal for virtual loudspeakers is binauralized to headphones.

The Ambisonics input signal, in this case FOA, is decomposed into a parametric DirAC representation [25]. This consists of a complex spectrum  $P(k, n)$ , where  $k$  denotes the frequency bin and  $n$  the time frame. For each time frequency bin, a diffuseness  $\psi$  and unit length direction vector  $\mathbf{r}_r$  are estimated. The sound scene is thereby decomposed into a diffuse and direct component, cf. [26]. The directed sound is complemented by distance information for each time frame. It is formulated as distance to the closest potential sound source in a spherical coordinate system. The mapping function  $m(\mathbf{r}, n)$  returns a distance in meters for each direction vector  $\mathbf{r}$  and time frame  $n$ .

The direction vector then undergoes different transformation steps. First, the distance according to the distance map is added by multiplying the unit direction vectors with the corresponding distance map entry:

$$\mathbf{d}_r(k, n) = \mathbf{r}_r(k, n) m(\mathbf{r}_r(k, n), n), \quad (1)$$

then the translation by the listener position  $\mathbf{l}(n) = [l_x(n), l_y(n), l_z(n)]^T$  is accounted for by subtracting it from each direction vector:

$$\mathbf{d}_l(k, n) = \mathbf{d}_r(k, n) - \mathbf{l}(n). \quad (2)$$

Additionally, the distance vector’s length is compensated to map the level change with respect to the closest source given by the distance map:

$$\mathbf{d}_v(k, n) = \frac{\mathbf{d}_l(k, n)}{\|\mathbf{d}_r(k, n)\|}. \quad (3)$$

The resulting distance vector  $\mathbf{d}_v(k, n)$  is then rotated according to the listeners orientation  $\mathbf{o}(n)$ . It can be written as vector composed of the pitch, yaw, and roll

$\mathbf{o}(n) = [o_x(n), o_z(n), o_y(n)]^T$ , which allows implementing the transformation using 2D rotation matrices, cf. eqn. (23) in [9]:

$$\mathbf{d}_p(k, n) = \mathbf{R}_Y(o_y(n))\mathbf{R}_Z(o_z(n))\mathbf{R}_X(o_x(n))\mathbf{d}_v(k, n). \quad (4)$$

The parametric representation is then decoded into virtual loudspeaker signals following an edge fading amplitude panning (EFAP) panning scheme [27] with the DirAC method. The angle of the unit vector  $\mathbf{r}_p$  is used for the panning, and the length of the vector  $\|\mathbf{d}_p\|$  is used for a distance dependent gain. The diffuse sound component is reproduced equally to all loudspeakers in order to provide an undirected ambience. Cf. [5] and references therein for more mathematical detail.

The channel signals are binauralized by convolving each virtual loudspeaker signal of the 8+4+4 setup with a HRTF for left and right ear. The distance of all speakers is fixed and no additional loudness change is added.

### 3. Experiments

The experiments follow a MUSHRA-like paradigm adopted for VR [24]. The different methods are compared and rated on a scale of 0 to 100 points, with 0 being the worst and 100 being perfect. There is a reference condition, which can be selected explicitly in addition. It is also one of the presented choices, this hidden reference is to be rated with 100 points. There is one clearly bad condition, the so called anchor. In order to do so in VR, the MUSHRA panel can be opened any time by the subject. They can then switch and rate the different renderings at will.

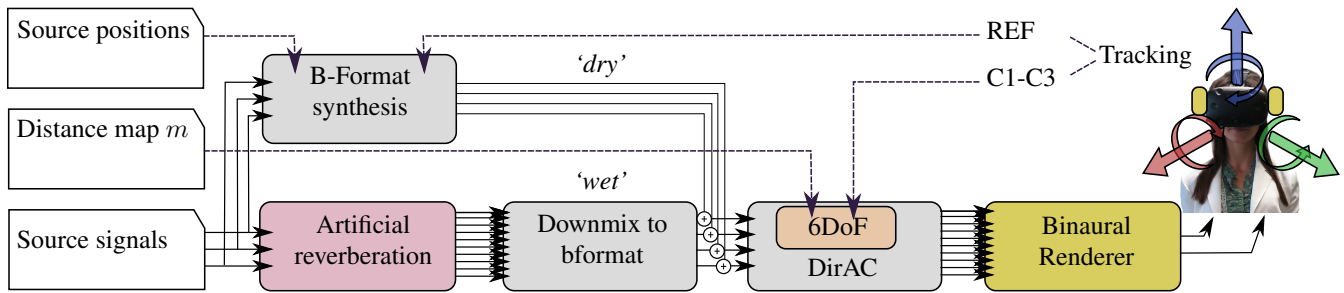
#### 3.1. Conditions

The four randomized conditions in our experiments were:

**REF** Object-based rendering. This is the reference condition. The B-format is generated on the fly for the listener’s current position and then rendered via the virtual loudspeakers.

**C1** 3DoF reproduction. The listener position is ignored, i.e.  $\mathbf{l}(n) = \mathbf{0}$ , but the head rotation  $\mathbf{o}(n)$  is still applied. The





**Fig. 4:** The signal paths for reference rendering and DirAC. In the reference case, the tracking data is used to change the positioning and rotation of the object-based synthesis (top left). In the other conditions C1-C3, the tracking data is applied in the DirAC domain (right).

gain is set to that of sources in a distance of 2 m from the listener. This condition is used as an anchor.

- C2** The proposed method for 6DoF reproduction without distance information. The listener position is used to change the direction vector. All sources are located on a sphere outside of the walking area. The radius of the sphere was fixed to 2 m and the distance-dependent gain is applied.
- C3** The proposed method of 6DoF reproduction with distance information. The listener position and orientation is accounted for. The distance information is used to compute the correct DOA at the listener position, and the distance-dependent gain is applied.

So the reference is not a perfect synthetic rendering, but the best possible recording position. This way the limitations of the FOA signal, namely the large width of the sources, is still audible. Nevertheless the spatial adaptation to the listeners movement is perfect. Similarly, the anchor is not a clearly bad low pass filtered version but just the 3DoF version. So it is spatially deficient in the regard that there is no distance attenuation or changing of DOAs depending on the tracked listeners' position but otherwise sounds plausible.

### 3.2. Technical realization

The experiment was realized using a HTC VIVE for tracking and reproduction, Unity 3D Engine for Graphics, and Max MSP for audio. The platform is described in full detail in [24]. The user wears the VIVE HMD and DT 770 Pro headphones driven by an RME Babyface, both connected to a PC running Unity and Max. A unity script encapsulates the tracking data and sends it as an open sound control (OSC) message to the Max patch. The switching of the active rendering as well as the MUSHRA rating and scene switching were realized by dedicated interaction scripts sending OSC messages to Max such as 'condition 2 selected', 'active condition rated 45', etc. Note that the Max patch took care of the randomization of conditions internally, so that neither the user nor the Unity scripts would know which is which.

In order to investigate the ability of the proposed method to reproduce the sound reproduced as if recorded at a single location in 6DoF, a dedicated rendering pipeline was constructed as shown in Fig. 4. A collection of virtual studio technology (VST) plugins was integrated using Max MSP 7. The key principle was to keep the same processing chain for all test conditions, so that the timbre is as similar as possible.

An FOA signal is generated from each source with distance attenuation with a dedicated VST in the Max patch. In case of the reference condition, the virtual microphone was placed at the listeners tracked position. In all other conditions, it was the fixed recording position in the center of the walking areas.

In scene 1, artificial reverberation is added to the source signal in a time-invariant manner by a dedicated VST (Fraunhofer IIS Reverb). Early reflections from the boundaries of the shoebox-shaped room are added with accurate delay, direction, and attenuation. Late reverberation is generated with a spatial feedback delay network (FDN) which distributes the multichannel output to the virtual loudspeaker setup [28]. The frequency-dependent reverberation time  $T_{60}$  was between 90 to 150 ms with a mean of 110 ms. A tonal correction filter with a lowpass characteristic was applied subsequently.

This is rendered to a 8+4+4 virtual speaker setup. Eight speakers are uniformly distributed along the medial plane at 0, 45, 90 degrees etc. An additional four are placed both at the top and bottom in a cross formation at  $\pm 45^\circ$  elevation. The reverberated signal is then converted to B-format by multiplying each of the virtual speaker signals with the B-format pattern of their DOA in an Ambisonics encoder VST. The reverberant B-format signal is added to the direct signal.

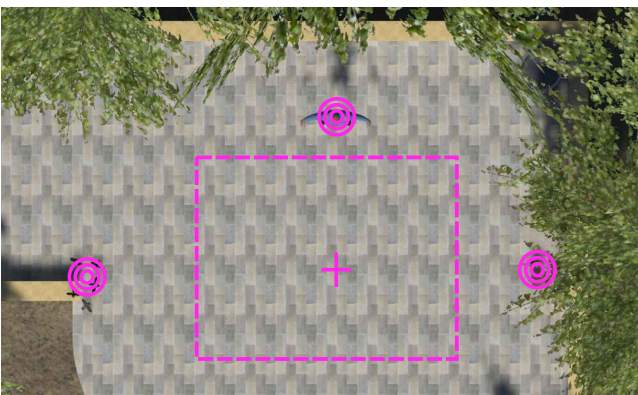
Subsequently, the mixed signal is processed in the parametric DirAC domain with a dedicated VST we developed. In case of the reference condition, no changes are made based on the tracking data and it is set to the recording position ( $\mathbf{l} = [0, 0, 0]^T$  and  $\mathbf{o} = [0, 0, 0]^T$ ). The processing is only done to keep the timbre and delays identical. An added benefit is that the switch between conditions can be realized seamlessly. In the other conditions, the 6DoF processing based on the tracking is applied to varying degree. In C3, only rotation is applied. In C2 and C3 translation according to the listener position is applied as well. Only in C3 the distance information is used. In C2, all sources are assumed slightly outside the walking area. The signal is then converted into a channel signal for a 8+4+4 loudspeaker configuration using EFAP panning [27].

These signals are the convolved with generic HRTFs with another VST. Each of the 16 channels is convolved with a far-field HRTF corresponding to the spherical coordinates for both left and right ear. Then the signal is output from Max to the headphones.





**Fig. 5:** The indoor scene. The sound is coming from the person, the radio and the open window, each source marked with concentric circles. The microphone position is marked by a cross. The user can walk in the area marked by the dashed rectangle on the floor.



**Fig. 6:** The outdoor scene. The sound is coming from the person, the radio and the birds, each source marked with concentric circles. The microphone position is marked by a cross. The user can walk in the area marked by the dashed rectangle on the floor.

### 3.3. Scenes

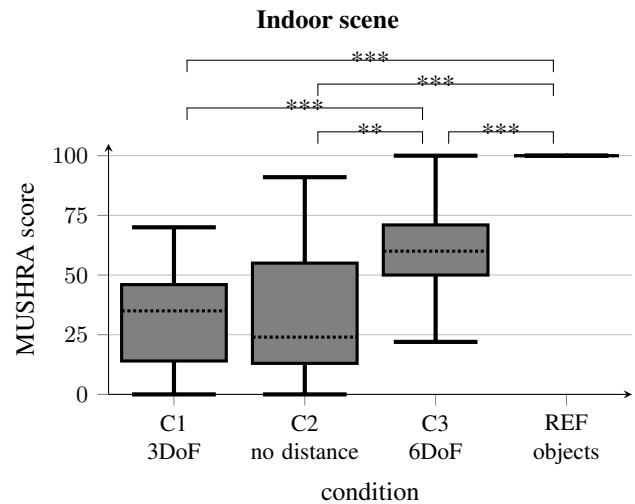
For the experiments, a room with an walking area of about  $3.5 \times 4.5\text{m}$  was used. The listening tests were conducted separately in two scenes. Both scenes were constructed realistically with visual representations of the sound objects. This was done to provide a baseline visual immersion for VR content.

The first scene was an indoor scene in a virtual living room. **Fig. 5** shows a top view. The cross shows the recording position. In about 0,5 m a virtual human speaker is placed, in 1 m distance a radio playing a string loop. In 2 m distance, just at the edge of the walking area, birds singing could be heard from a window.

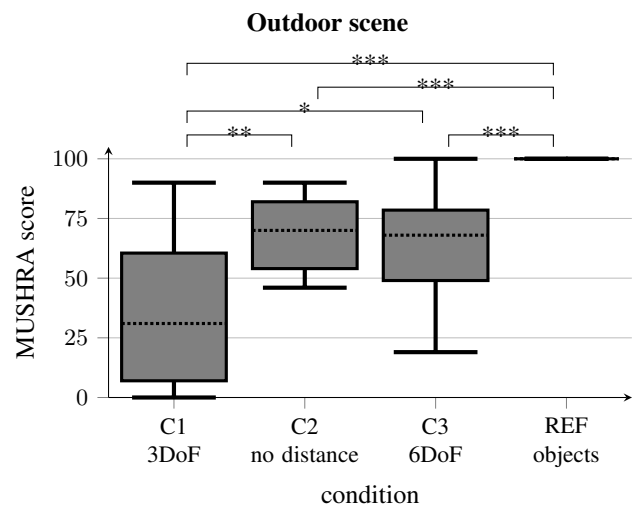
The second scene was placed outdoors. **Fig. 6** show a top view. The same sound sources were played back, but this time outside the walking area. The human was placed in 2 m, the radio in 3 m and the birds, this time in view as birds on the floor, in 4 m distance.

## 4. Results

The listening tests were conducted on different days with a total of 25 subjects, some only participating on the first



**Fig. 7:** MUSHRA ratings for the indoor scene (N=20).



**Fig. 8:** MUSHRA ratings for the outdoor scene (N=18).

day. They were 24-40 years old, male and female. Their VR experience had a wide range from almost none to quite extensive. After excluding subjects that scored the reference with less than 90 points, there are 20 remaining out of 24 for scene one and 18 out of 23 for scene two. The scores for both scenes are shown in **Fig. 7** and **Fig. 8** as box plots. The dotted line represents the median score, the boxes the 1st to 3rd quartile, the whiskers are at  $\pm 1.5$  inter-quartile range (IQR). Stars indicate significant differences according to a pairwise permutation test using one million permutations [29],  $* = p < 0.1$ ,  $** = p < 0.01$ ,  $*** = p < 0.001$ .

### 4.1. Indoor Scene

As indicated in **Fig. 7**, all 20 subjects scored the reference with 100 points in the indoor scene. There was no significant difference between C1 and C2 ( $p = 0.74$ ) that is both the 6DoF without distance and the 3DoF were rated worst with around 30 points. In both cases the sound came from the wrong side when walking behind sources, which may have dominated all other effects. The 6DoF scheme with distance information is rated better with around 60 points. The difference between using distance information (C3) or not

(C2) is significant ( $p \leq 0.01$ ). The reference is clearly valued better than all other conditions ( $p \leq 0.001$ ), as is 6DoF with distance (C3) vs 3DoF (C1) ( $p \leq 0.001$ ).

## 4.2. Outdoor Scene

As can be seen in Fig. 8, all 18 subjects scored the reference with 100 points in the outdoor scene. In contrast to the indoor scene, there was no significant difference between the 6DoF rendering with (C3) and without (C2) distance information ( $p = 0.70$ ). The 3DoF rendering (C1) was ranked lower but not as clearly. However it is rated significantly below both C3 ( $p \leq 0.02$ ) and C2 ( $p \leq 0.01$ ). Again, the reference is clearly valued higher than all other conditions ( $p \leq 0.001$ ).

## 4.3. Discussion

The 3DoF anchor is harder to spot than the usual anchors, which can confuse subjects, especially those used to non-VR MUSHRA tests. This can explain the rather big variance of scores. In, e.g., [30] mono mix was used as anchor, which leads to a clearer distinction. This could have been used in addition. Still, the requirement of a reference condition and a lower anchor condition is tricky in VR. Alternative methods are emerging to avoid this problem [31].

We did not exclude results when subjects rated C1 and C2 similar, which happened often in the first scene as the DOA was audibly wrong for both. There is a strong and significant distinction between cases with correct and faulty DOA in the indoor scene. This distinction is shifted towards a lower bar of 3DoF in the second scene. As long as there was a perceptible loudness change and reasonable DOAs, the scene was accepted as good. Even though there was a misplacement by using 2 m instead of the true up to 4 m distance of the sources, there was no significant difference in the subject's rating of C2 and C3.

## 5. Conclusion

A novel method for reproducing spatial audio recordings in 6DoF was evaluated. The method employs distance information to reproduce sound recorded at a single position at different points in the space the listener can move in. This distance information is important in scenarios with close sources, which the listener can move around. A listening test was conducted in two separate scenarios. The first was an indoor scenario with sources reachable by the listener; the second was an outdoor scenario where the sound sources are unreachable. A MUSHRA test showed a clear significant preference for the use of distance information in the first scenario, but not in the second. In the second scenario the 6DoF reproduction was clearly preferred to the 3DoF with no distance attenuation. This indicates an enhanced realism by correct placement for close sources and by distance attenuation for sound source in close to medium distance.

## References

[1] D. A. Bowman and R. P. McMahan, "Virtual reality: How much immersion is enough?," *IEEE Computer*, vol. 40, no. 7, pp. 36–43, 2007.

[2] H. Wallach, "The role of head movements and vestibular

and visual cues in sound localization.," *Journal of Experimental Psychology*, vol. 27, no. 4, pp. 339, 1940.

- [3] K. I. McAnally and R. L. Martin, "Sound localization with head movement: implications for 3-D audio displays," *Frontiers in Neuroscience*, vol. 8, pp. 210, 2014.
- [4] H. Hacihabiboglu, E. D. Sena, Z. Cvetković, J. Johnston, and J. O. Smith III, "Perceptual spatial audio recording, simulation, and rendering: An overview of spatial-audio techniques based on psychoacoustics," *IEEE Signal Processing Magazine*, vol. 34, no. 3, pp. 36–54, May 2017.
- [5] A. Plinge, S. J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, and E. A. P. Habets, "Six-degrees-of-freedom binaural audio reproduction of first-order Ambisonics with distance information," in *Audio Engineering Society International Conference on Audio for Virtual and Augmented Reality*, Redmond, U.S.A., Aug. 2018.
- [6] M. Frank, F. Zotter, and A. Sontacchi, "Producing 3D audio in Ambisonics," in *Audio Eng. Soc. Conf.*, Mar. 2015.
- [7] D. P. Jarrett, E. A. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*, Springer, 2017.
- [8] A. Plinge, F. Jacob, R. Haeb-Umbach, and G. A. Fink, "Acoustic microphone geometry calibration: An overview and experimental evaluation of state-of-the-art algorithms," *IEEE Signal Processing Magazine*, vol. 33, no. 4, pp. 14–29, July 2016.
- [9] M. Kronlachner and F. Zotter, "Spatial transformations for the enhancement of Ambisonic recordings," in *2nd International Conference on Spatial Audio*, Erlangen, Germany, 2014.
- [10] K. Kowalczyk, O. Thiergart, M. Taseska, G. Del Galdo, V. Pulkki, and E. A. P. Habets, "Parametric spatial sound processing: A flexible and efficient solution to sound scene acquisition, modification, and reproduction," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 31–42, 2015.
- [11] O. Thiergart, K. Kowalczyk, and E. A. P. Habets, "An acoustical zoom based on informed spatial filtering," in *International Workshop on Acoustic Signal Enhancement*, Sept. 2014, pp. 109–113.
- [12] H. Khaddour, J. Schimmel, and F. Rund, "A novel combined system of direction estimation and sound zooming of multiple speakers," *Radioengineering*, vol. 24, no. 2, June 2015.
- [13] O. Thiergart, G. D. Galdo, M. Taseska, and E. A. P. Habets, "Geometry-based spatial sound acquisition using distributed microphone arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2583–2594, 2013.

- [14] C. Schörkhuber, P. Hack, M. Zaunschirm, F. Zotter, and A. Sontacchi, "Localization of multiple acoustic sources with a distributed array of unsynchronized first-order Ambisonics microphones," in *Congress of Alps-Adria Acoustics Association*, Graz, Austria, Jan. 2014.
- [15] J. G. Tylka and E. Y. Choueiri, "Soundfield navigation using an array of higher-order Ambisonics microphone," in *Audio Eng. Soc. Conf. on Audio for Virtual and Augmented Reality*, Los Angeles, California, U.S.A., 2016.
- [16] E. Patricio, A. Ruminski, A. Kuklasinski, L. Januszkiewicz, and T. Zernicki, "Toward six degrees of freedom audio recording and playback using multiple Ambisonics sound fields," in *Audio Engineering Society Convention*, York, U.K., Mar. 2019.
- [17] J. Zamojski, P. Makaruk, L. Januszkiewicz, and T. Zernicki, "Recording, mixing and mastering of audio using a single microphone array and audio source separation algorithms," in *Audio Engineering Society Convention*, New York, U.S.A., Oct. 2017.
- [18] A. Iljazovic, F. Leschka, B. Neugebauer, and J. Plogsties, "The influence of 2-d and 3-d video playback on the perceived quality of spatial audio rendering for headphones," in *Audio Engineering Society Convention*, Oct. 2012.
- [19] S. Merchel and S. Groth, "Adaptively adjusting the stereophonic sweet spot to the listener's position," *Journal of the Audio Engineering Society*, vol. 58, no. 10, pp. 809–817, 2010.
- [20] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head related transfer functions of human subjects," *Journal of the Audio Engineering Society*, vol. 43, no. 5, pp. 300–321, 1995.
- [21] J. Blauert, *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*, The MIT Press, Oct. 1996.
- [22] M. Noisternig, A. Sontacchi, T. Musil, and R. Holdrich, "A 3D Ambisonic based binaural sound reproduction system," in *Audio Eng. Soc. Conf.*, June 2003.
- [23] Z. Ben-Hur, F. Brinkmann, J. Sheaffer, S. Weinzierl, and B. Rafaely, "Spectral equalization in binaural signals represented by order-truncated spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4087–4096, 2017.
- [24] T. Robotham, O. Rummukainen, J. Herre, and E. Habets, "Evaluation of binaural renderers in virtual reality environments: Platform and examples," in *Audio Engineering Society Convention*, New York, U.S.A., Oct. 2018.
- [25] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, vol. 55, no. 6, pp. 503–516, 2007.
- [26] O. Thiergart, G. Del Galdo, F. Kuech, and M. Prus, "Three-dimensional sound field analysis with directional audio coding based on signal adaptive parameter estimators," in *Audio Engineering Society Convention on Spatial Audio: Sense the Sound of Space*, Oct. 2010.
- [27] C. Borß, "A polygon-based panning method for 3D loudspeaker setups," in *Audio Engineering Society Convention*, Los Angeles, CA, U.S.A., Oct. 2014, pp. 343–352.
- [28] S. J. Schlecht and E. A. P. Habets, "Sign-agnostic matrix design for spatial artificial reverberation with feedback delay networks," in *Audio Engineering Society International Conference on Spatial Reproduction*, Tokyo, Japan, 2018.
- [29] P. Good, *Permutation Tests – A Practical Guide to Resampling Methods for Testing Hypotheses*, Springer Series in Statistics. Springer, 2 edition, 2000.
- [30] T. Robotham, O. Rummukainen, J. Herre, and E. A. P. Habets, "Online vs. offline multiple stimulus audio quality evaluation for virtual reality," in *Audio Engineering Society Convention*, New York, U.S.A., Oct. 2018.
- [31] O. Rummukainen, T. Robotham, S. J. Schlecht, A. Plinge, J. Herre, and E. A. P. Habets, "Audio quality evaluation in virtual reality: Multiple stimulus ranking with behavior tracking," in *Audio Engineering Society International Conference on Audio for Virtual and Augmented Reality*, Redmond, U.S.A., Aug. 2018.



## Full Reviewed Paper at ICSA 2019

Presented \* by VDT.

### Towards the Perception of Sound Source Directivity Inside Six-Degrees-of-Freedom Virtual Reality

Thomas Robotham<sup>1</sup>, Olli S. Rummukainen<sup>1</sup>, Emanuël A. P. Habets<sup>1</sup>

*International Audio Laboratories Erlangen<sup>1</sup>, Germany, email: thomas.robatham@iis.fraunhofer.de*

#### Abstract

Sound source directivity is a measure of the distribution of sound, propagating from a source object. It is an essential component of how we perceive acoustic environments, interactions and events. For six-degrees-of-freedom (6-DoF) virtual reality (VR), the combination of binaural audio and complete freedom of movement introduces new influencing elements into how we perceive source directivity. This preliminary study aims to explore if factors attributed to 6-DoF VR have an impact on the way we perceive changes of simple sound source directivity. The study is divided into two parts. Part I comprises of a control experiment in a non-VR monaural listening environment. The task is to ascertain difference limen between reference and test signals using a method of adjustment test. Based on the findings in Part I, Part II implements maximum attenuation thresholds on the same sound source directivity patterns using the same stimuli in 6-DoF VR. Results indicate that for critical steady-state signals, factors introduced by 6-DoF VR potentially mask our ability to detect loudness differences. Further analysis of the behavioral data acquired during Part II provides more insight into how subjects assess sound source directivity in 6-DoF VR.

#### 1. Introduction

The directivity of a source is a measure of the distribution of sound when propagated, dictated by its shape, size and material properties [12]. When sound is emitted into a diffuse-field environment, what arrives at our ears is a summation of direct and reflected components, all with individual characteristics, initially determined by the source's directivity pattern. For auralizations inside VR, realistic sound effects play an important role in our sense of presence [1] and for sound sources, the directivity is a key component [11]. Altering this directivity on the same audio source could potentially lead to manipulation of perceptual aspects such as localization [2], specifically distance [24], or even increased presence [25].

Some research regarding auralizations suggest that alterations in source directivity *can* be perceived by subjects [4], even

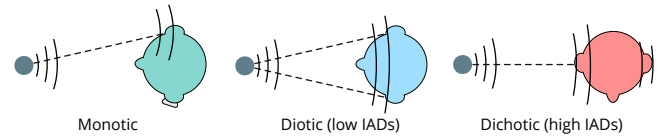
in complex free-field/outdoor environments [8]. However, a comparison between objective and subjective evaluations of source directivity suggests that whilst a significant difference in objective parameters is present between omni-directional and 'realistic' instrument sources, results of subjective testing showed no perceived significant difference [22]. Although not stated as significant, subjects were however able to perceive differences between omni-directional and a highly narrow-beamed source, with  $1/16^{th}$  of its surface area set to 10 dB louder than the remaining area. As the evaluation was auralized at various static positions, these results raise the question if such differences could be audible within a 6-DoF VR environment. A recent study conducted by Sloma and Neidhardt [18] explores such effects. Using two characteristics of directivity, omni-directional and loudspeaker data, various testing phases were conducted involving a static position with head movements, and full movement with guided paths. The aim was to state, for all phases, which source directivity was

<sup>1</sup>A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS).

present. Subjects should stick to pre-determined trajectories, which were: a straight line (provoking fluctuations in distance attenuation) and a portion of the circumference (maintaining a constant distance). The selected stimulus was music, additional room acoustics were present. The authors conclude by stating that subjects were unable to distinguish between sources with different directivity responses at static positions with head rotations, that room acoustics only have a high influence during static listening positions, and that listener movement itself has a considerable contribution to our perception of source directivities. However, one could speculate that the simultaneous presence of distance attenuation, non-steady-state signal type, and inclusion of room acoustics may heavily influenced the results.

Considering a source inside an ideal anechoic environment, we hear only the direct sound. If this source has a frequency-dependent directional response, walking around the sound source alters timbre and loudness due to magnitude changes in frequency spectrum. At discrete angular azimuth ( $\theta$ ) positions A and B ( $A \neq B$ ), the two frequency responses possess a difference. If the frequency response remains the same at points A and B but only overall gain changes (i.e., frequency independent directivity), we are essentially comparing the loudness at two different positions of the same stimuli. Difference limen (DL), that is the just noticeable difference (JND) of a given attribute, for loudness/intensity have been extensively researched. Depending on method, aural presentation (monaural, binaural etc.), measure, presentation level, and stimuli, the results may differ. However, some consensus within literature suggests for intensity discrimination using signals presented monotically (see Figure 1) at 40 dB SL, thresholds lie around  $\approx 0.5$  dB [9, 10], with subjects being less sensitive at levels below 40 dB SL [7, 15]. The unit dB SL refers to Sensation Level. This is the level in dB above individual listeners auditory threshold [23]. The effect remains similar when stimuli are presented diotically with subjects being slightly more sensitive [3] (i.e., we are able to detect smaller differences). For a detailed review of intensity DLs, we refer the reader to [17]. The problem becomes even more complicated with the introduction of varying interaural differences (IADs) or with non-steady-state signals [13].

The main aim of this preliminary study is to validate if the inclusion of 6-DoF elements directly influences our ability to perceive a loudness difference of  $A \neq B$  inside VR. This includes (but not limited to) two main aspects. One, the inclusion of consistent movement means an instant A/B comparison is not possible between discrete points. Therefore, the speed of movement, or rate of change may influence our ability to notice such differences. Two, as mentioned, the orientation of the listener's head, also influences noticeable level differences. Again, this may be further affected by consistent fluctuations in IAD states due to consistent head movements. The variation of loudness within this 6-DoF context is directly mapped to the subjects' position relative to the sound source and thus, its directivity. Further information may also be gained from behavioural analysis of tracking data, regarding how subjects conducted the listening tests. The scope is not to ascertain new DL for 6-DoF VR, but to



**Fig. 1:** Presentation of various head orientations in determining specific DL.

validate if deviations from traditional DL levels are present in the context of source directivity.

## 2. Study Overview

The study is divided into two experiments. Experiment I ascertains JND data from a control experiment with signals presented in a diotic manner (Figure 1). Experiment II explores if any factors associated with 6-DoF VR have any effect on detecting changes in loudness in the context of source directivity, based on data from Experiment I. Throughout Experiments I and II, both the directivity functions and stimuli remain constant.

### 2.1. Directivity Patterns

Two first-order directivity patterns were chosen for investigation: A) cardioid and B) dipole. Both patterns are two of the most commonly used for source directivity, as they are simple to describe using a parametric function. So called zero- and first-order directivity patterns (such as omnidirectional, cardioid, bi-directional, super-cardioid, etc.) can be constructed by computing a weighted sum of an omnidirectional and a dipole pattern. As such, they are integrated into most 6-DoF audio renderers that include sound source directivity [6, 21]. Both cardioid and dipole patterns provide a smooth curve which subjects can easily understand as louder and quieter when walking in a specific direction around the sound source (Figure 2). The correlation between off-axis source and listener angle ( $\theta$ ) attenuation for the two patterns, are taken from [14, 19] and in terms of dB sound pressure level (SPL) are given as:

$$y \text{ dB (SPL)} = 20 \log_{10} \left( \frac{1}{2} \cos \left( \frac{\theta\pi}{180} \right) + \frac{1}{2} \right) \quad (1)$$

$$y \text{ dB (SPL)} = 20 \log_{10} \left( \cos \left( \frac{\theta\pi}{180} \right) \right), \quad (2)$$

where  $\theta$  is the angle in degrees. The relationship between the two patterns also provides further information. As  $\theta$  increases from  $0^\circ$ , the angular distance which the user must 'move' around a dipole pattern to reach equal levels of attenuation from a cardioid pattern is halved (illustrated in Figure 2). Therefore, a comparison of results between both known rates of attenuation may provide further insight into subjects ability to detect changes, without them having to physically 'move' twice as fast.

The scope of this study limits the source directivity patterns to be frequency independent. However, due to the equal



loudness contour of the human ear, it is possible that the results for a broadband attenuation are influenced by certain frequency bands more than others.

## 2.2. Stimuli

Depending on the psychophysical methodology employed, steady-state signals are often presented to subjects to eliminate the effects of level fluctuation [13]. As such, pink noise was selected as a steady-state signal allowing subjects to be highly critical with no temporal variations. However, in 6-DoF VR, there are seldom situations in which steady-state signals are present, either due to the signals themselves (music, speech, environmental, etc.) or the influence of distance attenuation provided by the subjects' movements. Therefore, two non-steady-state signals were chosen, viz. anechoic male speech and anechoic cello. Using a CORTEX head and torso simulator and Bayerdynamic DT770 closed headphones, playback was calibrated to an absolute level of 67 dB SPL using pink noise. Any louder than this, playback would become uncomfortable for louder portions of the non-steady-state signals. Normally, subjects' individual auditory thresholds should also be added to the playback level however, due to limited availability, individual auditory thresholds were not measured and accounted for, and the reproduction system was calibrated to an absolute playback level. It should be noted that this procedure can result in larger deviations across subjects.

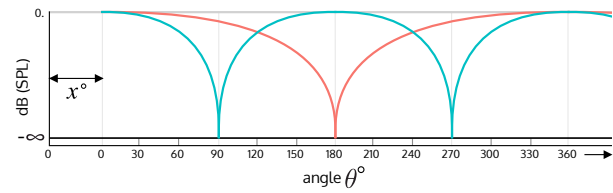
## 3. Experiment I

The goal of Experiment I is to ascertain JND values of loudness for signals presented monaurally in terms of dB SPL using two attenuation functions.

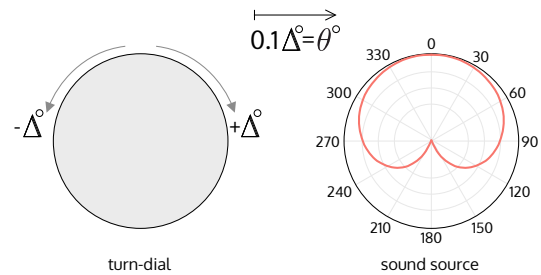
### 3.1. Method

The Method of Adjustment (MoA) was selected as the method to identify difference limen between a *reference* signal and a *test* signal [20]. The objective is to vary the level of the test signal, such that a JND in loudness can be heard when compared to the reference signal. Subjects may alternate between the reference and test signals at any time, with no limitation on the number of times they may compare. Both *reference* and *test* signals are presented monaurally over headphones where only one signal is played at a time. Whilst other psychophysical methods such as method of constant stimuli may be more accurate [26], MoA puts the test signal under the subjects control. The benefit of this approach is that the subject is an active contributor to finding the criteria, thus paying more attention, as opposed to being presented signals in a passive manner and asked to make a forced choice between two signals. This methodology is also more comparable to a 6-DoF VR scenario where, given a static source, the subject is able to indirectly influence loudness via their body movements (e.g., via distance attenuation or directivity) and, the stimulus presentation is continuous.

For the test, subjects controlled the test signal via a turn-dial button connected to a MaxMSP patch. The starting loudness



**Fig. 2:** Directivity patterns cardioid (orange) and dipole (blue) as a function of the angle  $\theta$ , where  $x$  represents the initial randomized angular distance.



**Fig. 3:** Mapping of the turn dial increments and decrements (left) to angular position around cardioid pattern (right).

of the both signals corresponds to the on-axis playback level at  $0^\circ$ . Rotating the dial resulted in a change in angular distance from the on-axis position; clockwise increases the angle, anti-clockwise decreases the angle. To eliminate any learning effects, the amount the dial must be initially rotated before the angular distance starts to increment is randomized across all items between  $0^\circ$  and  $30^\circ$  (see Figure 2). As the rotation of the dial emulates the movement around a sound source, continued clockwise movement results in the directivity pattern repeating itself after  $360^\circ$ . Finally,  $1^\circ$  rotation of the turn dial being equal to  $1^\circ$  angular increase along the directivity function was considered too sensitive. Therefore, it was remapped such that  $10^\circ$  rotation of the turn dial resulted in  $1^\circ$  angular change around the source (see Figure 3).

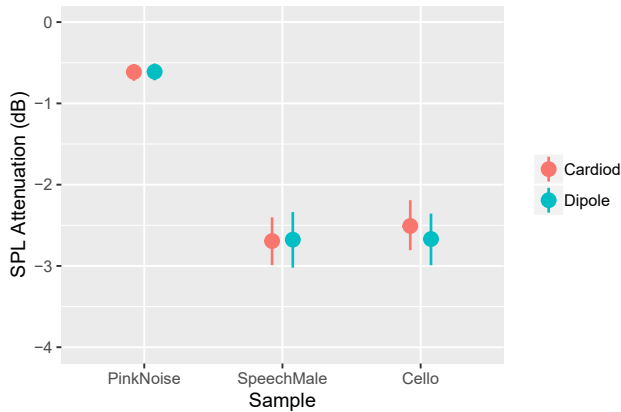
### 3.2. Subjects and Procedure

Twenty subjects participated in the control test, fifteen male and five female ranging from ages 20 - 42. All subjects were a mixture of trained and expert listeners at Fraunhofer IIS, none of which reported any hearing impairments. The test was conducted in a soundproof listening booth and instructions for subjects were presented both verbally and through text. Operation of the test was via keyboard used for switching between test and reference signals, and a turn-dial for the loudness control. After finding a JND value, subjects would press the turn-dial to move onto the next item. Each stimuli was repeated three times, resulting in  $(2_{Directivities} \times 3_{Stimuli} \times 3_{Rep})$  18 items. Average test time was around 20 minutes.

### 3.3. Results

The mean and 95% confidence interval (CI) of subjects' results for the control test can be seen in Figure 4. Statistical analysis was performed using an analysis of variance. Mean responses per sample show no significant difference between





**Fig. 4:** Mean JND values as a physical measure in dB (SPL) between reference and test signals. Whiskers denote the bootstrapped 95% confidence intervals.

directivity patterns. It is therefore reasonable to argue the rate of attenuation per angular turn of the dial, provided no bias into subjects response. A significant effect was found under the sample type used ( $F_{(2,38)} = 91.534, p < 0.001, \eta_G^2 = 0.532$ ). For the steady-state pink noise signal, mean values are  $\approx -0.6$  dB SPL in accordance with literature [17]. For non-steady-state signals, mean values are  $\approx -2.6$  dB SPL. Additionally, whilst the task proved harder for temporally fluctuating signals, CIs are still small and consistent across all samples and directivity patterns, suggesting that absolute level calibration at 67 dB SPL was sufficient for this test.

## 4. Experiment II

The goal of Experiment II, is to investigate if the introduction of self-motion and binaurally presented signals have an influence on subjects’ ability to detect the thresholds found in Experiment I when incorporated into two sound source directivity patterns.

### 4.1. Method

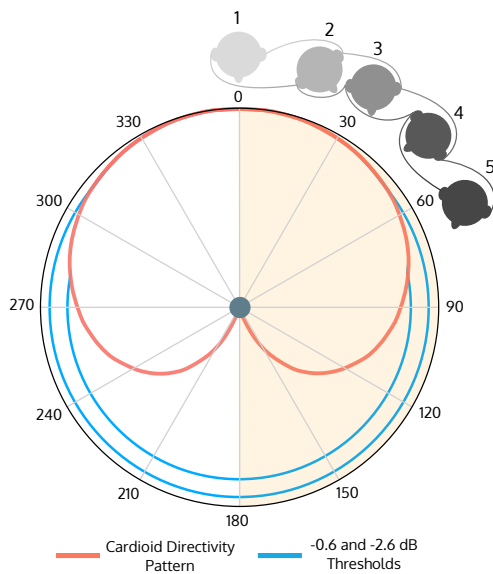
Inside 6-DoF VR, continuous movement allows the subject to fully explore around a sound source. As such, an instant A/B comparison with the subjects’ current position, against the on-axis position is not possible without by-passing the effect of self-movement. Therefore, to investigate whether continuous self-movement and constantly fluctuating binaural cues affect subjects’ ability to detect loudness change, a different test methodology is used. Values found in Experiment I for steady and non-steady state signals are employed as thresholds, limiting the attenuation of the directivity pattern at a certain level. As the subject walks around the sound source, the level changes according to the cardioid or dipole directivity pattern as expected, until the attenuation is equal to that of JND values found in Experiment I. At this point, the level is maintained until returning above the threshold (see Figure 5). If during the assessment the subjects are unable to detect a JND in signal level, it may be hypothesized that influencing factors involved in 6-DoF VR impede, or mask, the ability to detect the changes in loudness heard in Experiment I.

The limitation of this methodology is that it only provides a binary ‘Yes/No’ response, and further investigation would be required to ascertain if these JND values are either higher or lower. However, it is possible to extract data by monitoring the subject’s behavior within the 6-DoF VR environment. By additionally recording user behavior during each test item, factors such as; rate of change with respect to subject movements, and speed of head movements may be analyzed and cross-referenced with subjects ‘Yes/No’ responses.

To conduct Experiment II, a 6-DoF VR environment was created in Unity. The overall system architecture is described in more detail in [16]. For indicating the position of the sound source, a sphere was placed in the center of the VR world. The height of the sphere was tethered to the HMD height restricting the users movements explicitly around the lateral plane of the source. A path was rendered at a 1 meter radius around the source as a guide for subjects to walk along. The angle  $\theta$  between on-axis position ( $0^\circ$ ) and the users’ position was directly mapped to the cardioid and dipole directivity patterns. For binaural audio, a parametric renderer was integrated with interaural time and level differences modelled from a spherical head model [5]. No distance attenuation was modeled to ensure that level fluctuations were induced purely by the directivity pattern. Using a CORTEX head and torso simulator and Bayerdynamic DT770 closed headphones, the playback level was calibrated such that the pink noise, when presented diotically at the on-axis position, was 67 dB SPL (consistent with Experiment I). In addition to the thresholds ascertained in Experiment I (see Figure 5), two sanity check conditions were also added. One included a threshold at  $-10$  dB SPL, and the other with no attenuation at all. If the same changes always are audible by subjects, it is highly likely this leads to listener fatigue and reduced concentration levels, thus these sanity check conditions provide noticeable random variation and post-test subject screening. Finally, to remove any learning effects, an initial angular distance randomized between  $0^\circ$  and  $+30^\circ$  degrees must be walked before the attenuation curve begins.

### 4.2. Subjects and Procedure

The same twenty subjects who participated in Experiment I also participated in Experiment II. The test was conducted in a virtual reality lab at Fraunhofer IIS using the HTC Vive Pro system, of which the VR space was calibrated to a size  $2.3\text{ m} \times 2.0\text{ m}$ . Instructions for the subjects were to walk to a starting position and click thumbpad on the Vive controller. Then, a test item began playing and subjects should walk along a circular path exploring a  $180^\circ$  area of interest around the sphere (see Figure 5). The task was to answer if they could hear a JND in absolute playback level of the signal when exploring around the sound source. To answer, rotating the hand-held Vive controller  $\geq 30^\circ$  to the *left* and pressing the trigger means yes,  $\geq 30^\circ$  to the *right* and clicking means no. For visual feedback, the controller would turn green and orange, indicating the respective choices. After an answer was given, subjects would go back to the start place to repeat the process. Once the test was completed, a text prompt would appear informing subjects they had finished. Instructions for



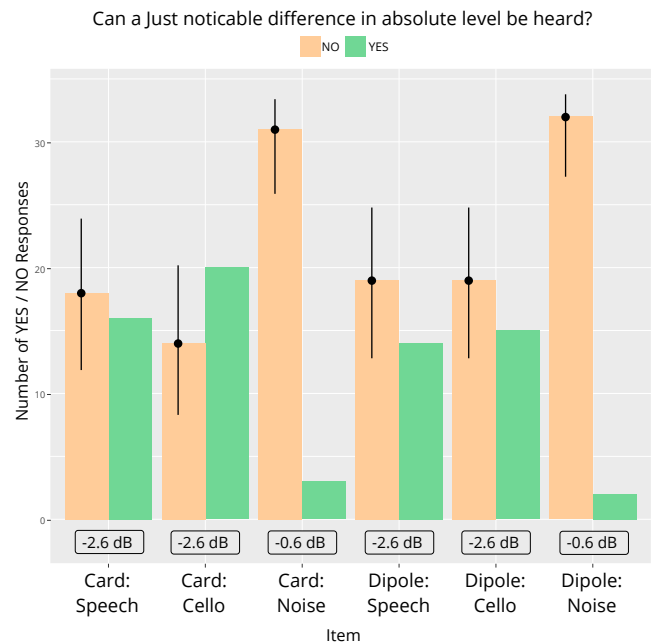
**Fig. 5:** Representation of  $-10$  dB SPL threshold for cardioid directivity pattern of sound source, with example of user walking around the area of interest (light orange).

the test were presented both verbally and also written inside the VR world for reference. Here, test items were only repeated twice, plus  $-10$  and  $0$  dB SPL sanity checks per stimuli resulting in a total of 21 items.

### 4.3. Results

The results of subjects' 'Yes/No' responses for Experiment II are shown in Figure 6. Green bars indicate the number of times subjects could hear a just noticeable difference, and orange if no difference in loudness could be heard. All subjects responded correctly to the sanity check thresholds ( $0$  dB = *could not* hear difference,  $-10$  dB = *could* hear a difference) and therefore, these responses are not plotted. Respective thresholds for the different stimuli are shown below each item. Binomial distribution tests were conducted for statistical analysis regarding the 'Yes/No' pairing of each item, and are overlaid on top of all 'No' responses. If the number of 'Yes' responses occupies the same range as the binomial test confidence intervals, the null hypothesis may be accepted (i.e., that no significant effect is present and subjects are equally likely to respond 'Yes' or 'No'). For non-steady-state signals Cello and Speech, no significant difference can be observed between subjects responding 'Yes/No'. However, for the noise signal, in almost all observations, subjects could *not* detect a difference in loudness. The number of 'Yes' responses does not overlap with the binomial test intervals therefore, a significant effect is present that results in subjects not being able to hear any difference. Both observations regarding the signal type can be made also for cardioid and dipole directivity patterns.

Comparing these results with the data from Experiment I, it is reasonable to conclude that for the critical noise signal, the inclusion of 6-DoF elements impacted subjects ability to notice any change in loudness. Mean values in Figure 4 indicate that 50% of subjects *could* detect a JND smaller than  $-0.6$  dB.



**Fig. 6:** Frequency of subjects' responses for Experiment II as bar plots. Binomial test data provided over the top showing probability of success for selected answer within 95% bootstrapped confidence intervals.

However, the results in Figure 6 show that a JND in loudness was detected only five times across all presentations. As this signal is steady state and no inherent temporal fluctuations are present, subjects' ability to recognize loudness changes may be hindered due to the shifting inter-aural time and level differences induced by head movements. The result is that this level difference of  $-0.6$  dB is no longer audible inside 6-DoF. Conversely, 'Yes/No' responses in Experiment II for the Cello and Speech signals remained statistically equal. As the results from Experiment I show that 50% of subjects could hear this difference, this may indicate that JND value of  $-2.6$  dB for non-steady-state signals has remained the same. As the methodology of Experiment II was designed to confirm or deny if subjects could hear previous findings, further testing would be needed to confirm this hypothesis. However, if such a threshold exists for non-steady state signals inside 6-DoF VR, this information could prove useful in perceptually optimizing directivity data for such sound sources.

## 5. Behavioral Analysis

### 5.1. Position, Orientation and Time

To further assess subjective results, tracking data of head movements and user position over time was recorded for each subject per test item. Analysis was conducted on total average: distance walked, head movements (pitch, yaw, roll and source relative yaw), and time taken per test item. For 6-DoF VR, exploring if freedom of movement combined with subjects' head rotations have any affect on our perception of source directivity is of particular interest. All analysis showed no significant difference between either the signal type used or the directivity pattern employed. In comparison

to the ‘Yes/No’ answers in Figure 6, where a large difference can be seen between steady and non-steady state signals, no such difference can be observed in the tracking data. This indicates ‘Yes/No’ responses provided by subjects were not due to uncertainty, which would be reflected in subjects needing significantly more time, or moving greater distances. Considering the relative yaw movement data, further insight may be gained by analyzing *how* subjects listened throughout the test.

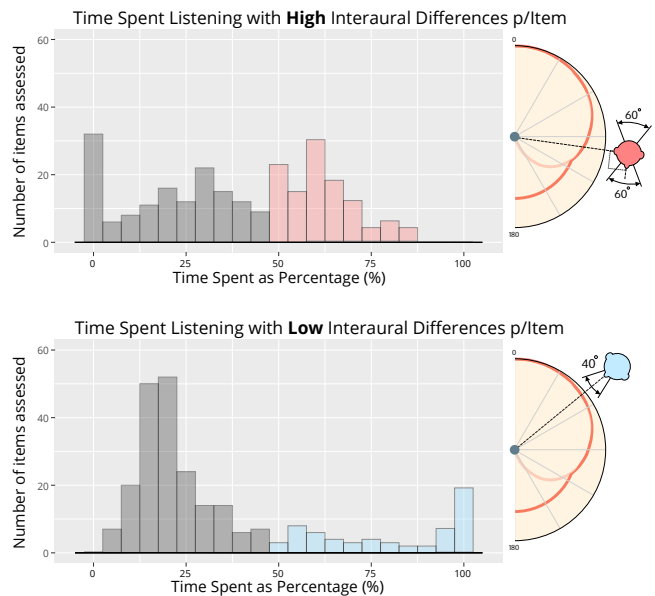
### 5.2. Interaural Differences

To assess if subjects listening with ‘high’ or ‘low’ IADs (orientations shown in Figure 1) had an effect on the results, source relative yaw tracking data was divided into ‘low’ and ‘high’ angular categories. For items that were assessed with relative head rotations (Yaw) of 0°, with ±20° variation were labeled as having ‘low’ IADs. Conversely, items that were assessed with 90° or 270° with ±30° variation were labeled as having ‘high’ IADs. The angular ranges (±20° and ±30°) for these categories was based on visual inspection of all raw tracking data. A smaller angular distribution was chosen for ‘low’ IADs as it appears easier for subjects to maintain a more accurate head position when looking at the source. For each subject, the percentage of time spent listening to each item with both IAD categories was calculated. The number of items where subjects spent over 50% of their time evaluating with ‘high’ or ‘low’ IADs was then counted and cross-referenced with their ‘Yes/No’ response. A visual representation of this IAD categorization is shown in Figure 7 and the results of cross-referencing responses with IADs are shown in Figure 8. This analysis aims to provide a preliminary insight into possible behaviors and not to establish new methods of analysis. By initial observation, the results in Figure 7 would indicate that *if* subjects were to listen for any JND whilst moving around a sound source in an intentional manner, the most frequent method would be with higher levels of IADs (as indicated in ‘Red’). Far fewer evaluations were conducted with subjects spending over 50% of their time with lower IADs.

For statistical analysis of data, a logistic regression model was used to determine if the categorized ‘high’ and ‘low’ IADs had a significant influence on subject responses. Generally, this method is used to determine if predictor variables (which may be both continuous or binary) can be used to model the log odds of a certain binary outcome (‘Yes/NO’). For this analysis, the predictor variables; ‘Directivity’, ‘Stimuli’, ‘Percentage of time spent with high IADs’ and ‘Percentage of time spent with low IADs’, were used to identify any significant effects on the binary ‘Yes/No’ outcome. Results in Table 1 show which of the variables have a significant effect. For categorical variables ‘Directivity’ and ‘Signal’, ‘Estimate’ shows the log odds of one variable over another changing the binary outcome. For continuous variables of IAD, every unit increase results in the log odds increasing by the estimate amount (i.e., the estimate equates to a single unit, hence why these are smaller values). From this, we can see the most significant effects on binary outcomes are the ‘Noise’ stimuli (already apparent in Figure 6), and when no attenuation curve

**Tab. 1:** Table of logit regression analysis showing significant predictive variables on the binary outcome of subjects being able to hear a JND in loudness.

Coefficients	Estimate	St. Err	Z-Val	P-value
<i>Intercept*</i>	-1.846	0.853	-2.164	0.031
Dipole × Cardioid	0.477	0.324	1.469	0.142
Dipole × Null***	-2.621	0.769	-3.406	0.001
Cello × Speech	-0.442	0.336	-1.257	0.209
Cello × Noise***	-2.798	0.527	-5.306	1.12e-07
High IADs	0.022	0.012	1.779	0.075
Low IADs*	0.025	0.010	2.431	0.015

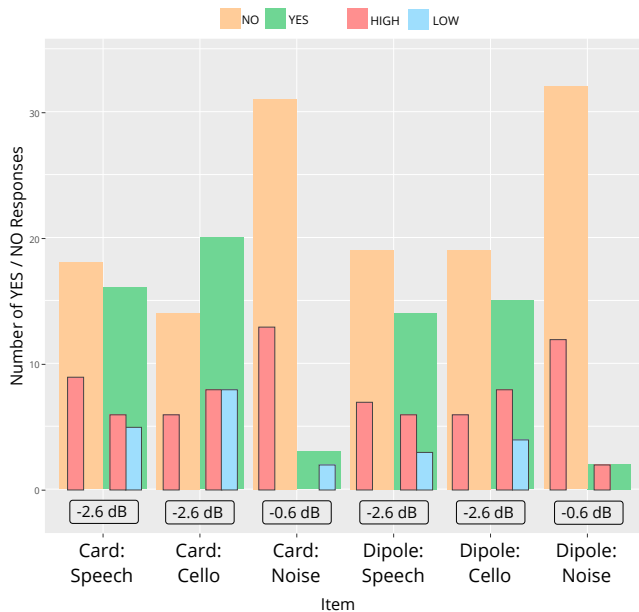


**Fig. 7:** Number of answers given against the time spent listening with specific IADs. Responses given where over 50% of time was spent listening with high IADs is shown in Red, and low IADs in Blue.

(‘Null’) is used. However, a significant effect is also found if subjects spent more than 50% of their time evaluating with low IADs. This result is also reflected in Figure 8 where *all* responses with ‘low’ IADs (blue) are in the ‘Yes’ column, suggesting we are more likely to hear a loudness difference when maintaining low IADs.

### 6. Discussion

For Experiment II, subjects were free to move around the sound source, no instructions were given advising subjects to maintain a specific head orientation. Given this, it is interesting to see that whilst conducting the evaluation with lower IADs made a significant impact on the outcome, it was far more prevalent to conduct the experiment with higher IADs. Furthermore, as previously mentioned, all responses given where subjects spent more than 50% of time listening with lower IADs, were *all* positive. This would imply that even though critical listening is more acute with consistent lower IADs, in 6-DoF VR where, subjects must move around the sound source, maintaining persistent head orientation towards the sound source is either too unnatural, or most subjects felt like they could conduct the task well in another



**Fig. 8:** Number of items listened to with high or low binaural cues overlaid on top of total ‘Yes/No’ responses. High binaural cues are highlighted in ‘Red’ and low in ‘Blue’.

manner. It is also highly unlikely that in a commercial 6-DoF VR scene there is only one audio source, and that users would walk round it intentionally maintaining low IADs. This raises an interesting question as to whether the same results are obtained if the user remains stationary and the directivity pattern changes due to the user rotating the sound source. This would be similar to Experiment I however, the signal would still be presented binaurally, and subjects are still free to move their heads. If results change such that significantly more subjects could hear a JND, this would provide further evidence to suggest natural body movements *combined* with higher interaural differences makes subjects less sensitive to changes in loudness inside 6-DoF VR. If our perception is effected in such a way, this may help in defining a more perceptually motivated model towards sound source directivity for 6-DoF VR. Furthermore, the scope of this investigation was limited to frequency independent attenuation. Due to interaural time and level differences operating at different frequency ranges, loudness variations in specific frequency bands may be more/less noticeable depending on the source relative head orientation of the subject.

## 7. Conclusion

Two experiments were conducted to investigate if the inclusion of body movements or binaural cues have an influence on our ability to detect broadband changes in loudness, relative to sound source directivity. Using the Method of Adjustment, Experiment I confirmed JND thresholds in literature for stimuli presented equally at both ears. The difference between steady and non-steady state signals was significant, however no difference was observed between the two attenuation functions (based on cardioid and dipole directivity patterns). Experiment II implemented the same

directivity patterns inside 6-DoF VR with attenuation limited to the thresholds found in Experiment I for respective stimuli. Subjects were asked to explore the area around the sound source and answer if a difference in absolute loudness could be heard. Results indicate that for the steady-state signal ‘Noise’, the inclusion of binaural cues and body movements meant that the JND of  $-0.6$  dB presented monaurally in Experiment I was no longer audible and that this threshold is higher inside 6-DoF VR. For both experiments, no significant difference was observed between the two directivity patterns. This may be due to the differences not being large enough over a given time and angular distribution. For future work, the authors aim to investigate various rates of change over angular distances to gain further insight into perceptual thresholds of source directivity inside 6-DoF VR. Finally, cross-correlating subject’s head movements with ‘Yes/No’ responses, showed that spending over 50% of time evaluating the sound source with lower IADs allowed subjects to always hear a just noticeable difference in loudness. This implies that our head orientation with respect to the sound source, and consequently varying IAD, does have effect on how we perceive sound source directivity inside 6-DoF VR. For future perceptual evaluations, this may be an important consideration depending on the task.

## 8. References

- [1] BEGAULT, D. R. *3D-Sound for Virtual Reality and Multimedia*. Academic Press, Boston, 1994.
- [2] BLAUERT, J. *Spatial Hearing: The Psychophysics of Human Sound Localization*, 2nd ed. MIT Press, London, England, 1997.
- [3] CHURCHER, B. G., KING, A. J., AND DAVIES, H. The Minimum Perceptible Change of Intensity of a Pure Tone. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 18, 122 (Nov. 1934), pp. 927–939.
- [4] DALENBÄCK, B.-I., KLEINER, M., AND SVENSSON, P. Audibility of Changes in Geometric Shape, Source Directivity, and Absorptive Treatment-Experiments in Auralization. *Journal of the Audio Engineering Society* 41, 11 (Nov. 1993), pp. 905–913.
- [5] DUDA, R. O., AND MARTENS, W. L. Range Dependence of the Response of a Spherical Head Model. *The Journal of the Acoustical Society of America* 104, 5 (1998), pp. 3048–3058.
- [6] GOOGLE INC. Resonance Audio Unity SDK API Reference, 2018. <https://valvesoftware.github.io/steam-audio/>
- [7] HIRSH, I. J., PALVA, T., AND GOODMAN, A. Difference Limen and Recruitment. *The Journal of the American Medical Association: Otolaryngology* 60, 5 (Nov. 1954), pp. 525–540.

- [8] HOARE, S., SOUTHERN, A., AND MURPHY, D. Study of the Effect of Source Directivity on the Perception of Sound in a Virtual Free-field. In *AES 128th Convention* (London, UK, May 2010), pp. 1–8.
- [9] KNUDSEN, V. O. The Sensibility of the Ear to Small Differences of Intensity and Frequency. *The American Physical Society: Physical Review* 21, 1 (Jan. 1923), pp. 84–102.
- [10] LÜSCHER, E., AND ZWISLOCKI, J. Comparison of the Various Methods Employed in the Determination of the Recruitment Phenomenon. *The Journal of Laryngology & Otology* 65, 03 (Mar. 1951), pp. 187–195.
- [11] MEHRA, R., ANTANI, L., KIM, S., AND MANOCHA, D. Source and Listener Directivity for Interactive Wave-Based Sound Propagation. *IEEE Transactions on Visualization and Computer Graphics* 20, 4 (Apr. 2014), pp. 495–503.
- [12] MEHRA, R., AND MANOCHA, D. Wave-based Sound Propagation for VR Applications. In *Proc. of IEEE VR Workshop: Sonic Interaction in Virtual Environments (SIVE)* (MN, USA, Mar. 2014), pp. 41–46.
- [13] NAMBA, S., KUWANO, S., AND FASTL, H. Loudness of Non-Steady-State Sounds. *Japanese Psychological Research* 50, 4 (2008), pp. 154–166.
- [14] PETERS, N., AND SCHMEDER, W., A. Beamforming Using a Spherical Microphone Array Based on Legacy Microphone Characteristics. In *Proc. of International Conference on Spatial Audio* (Detmold, Germany, 2011), pp. 1–7.
- [15] RIESZ, R. R. Differential Intensity Sensitivity of the Ear for Pure Tones. *The American Physical Society: Physical Review* 31, 5 (May 1928), pp. 867–875.
- [16] ROBOTHAM, T., RUMMUKAINEN, O., HERRE, J., AND HABETS, E. A. P. Evaluation of Binaural Renderers in Virtual Reality Environments: Platform and Examples. In *AES 145th Convention* (NY, USA, Oct. 2018), pp. 1–5.
- [17] ROWLAND, R. C., AND TOBIAS, J. V. Interaural Intensity Difference Limen. *Journal of Speech, Language, and Hearing Research* 10, 4 (Dec. 1967), pp. 745–756.
- [18] SLOMA, U., AND NEIDHARDT, A. Investigations on the Impact of Listener Movement on the Perception of Source Directivity in Virtual Acoustic Environments. In *Proc. of Annual German Conference on Acoustics* (Munich, Germany, 2018), pp. 1–4.
- [19] SOUTHERN, A., AND MURPHY, D. Low Complexity Directional Sound Sources for Finite Difference Time Domain Room Acoustic Models. In *AES 126th Convention* (Munich, Germany, May 2009), pp. 1–10.
- [20] STEVENS, S. S. The Measurement of Loudness. *The Journal of the Acoustical Society of America* 27, 5 (Sept. 1955), pp. 815–829.
- [21] VALVE. Steam Audio, 2019. <https://valvesoftware.github.io/steam-audio/>
- [22] WANG, L. M., AND VIGEANT, M. C. Objective and Subjective Evaluation of the Use of Directional Sound Sources in Auralizations. Tech. rep., University of Nebraska, Nebraska, Lincoln, (Apr. 2004), pp. 2711-2714
- [23] WARD, W. D. Use of Sensation Level in Measurements of Loudness and of Temporary Threshold Shifts. *The Journal of the Acoustical Society of America* 39, 4 (Apr. 1966), pp. 736–740.
- [24] WENDT, F., FRANK, M., ZOTTER, F., HÖLDRICH, R. Directivity Patterns Controlling the Auditory Source Distance. In *Proc. of 19th International Conference on Digital Audio Effects* (Brno, Czech Republic, 2016), pp. 295–300.
- [25] WENDT, J., WEYERS, B., AND VIERJAHN, T. Does the Directivity of a Virtual Agent’s Speech Influence the Perceived Social Presence? In *Proc. of IEEE Virtual Humans in Crowds for Immersive Environments (VHCIE)* (Reutlingen, Germany, Mar. 2018), pp. 1–3.
- [26] WIER, C. C., JESTEADT, W., AND GREEN, D. M. A Comparison of Method-of-Adjustment and Forced-Choice Procedures in Frequency Discrimination. *Perception & Psychophysics* 19, 1 (Jan. 1976), pp. 75–79.





# Abstract Reviewed Paper at ICSA 2019

Presented \* by VDT.

## Deep Neural Network Approaches for Selective Hearing based on Spatial Data Simulation

S. Hestermann, H. Lukashevich, C. Sladeczek  
*Fraunhofer Institute for Digital Media Technology IDMT*

### Abstract

Selective Hearing (SH) refers to the listener's attention to specific sound sources of interest in their auditory scene. Achieving SH through computational means involves detection, classification, separation, localization and enhancement of sound sources. Deep neural networks (DNNs) have been shown to perform these tasks in a robust and time-efficient manner. A promising application of SH are intelligent noise-cancelling headphones, where sound sources of interest, such as warning signals, sirens or speech, are extracted from a given auditory scene and conveyed to the user, whilst the rest of the auditory scene remains inaudible. For this purpose, existing noise cancellation approaches need to be combined with machine learning techniques. In this context, we evaluate a convolutional neural network (CNN) architecture and a long short-term memory (LSTM) architecture for the detection and separation of sirens. In addition, we propose a data simulation approach for generating different sound environments for a virtual pair of headphone microphones. The Fraunhofer SpatialSound Wave technology is used for a realistic evaluation of the trained models. For the evaluation, a three-dimensional acoustic scene is simulated via the object-based audio approach.

### 1. Introduction

Conventional closed-back headphones block environmental sounds through insulated ear cups. Noise-cancelling headphones, on the other hand, use on-board processing to cancel ambient sounds through destructive interference [8]. Under certain conditions, this technology poses one significant problem. Since most algorithms rely on basic physical principles, they lack semantic understanding of the canceled signal. Consequently, any sound is blocked, regardless of its potential importance to the headphone user. Information- and time-critical sounds, such as sirens in traffic, thus may not receive the user's attention and provoke dangerous situations.

This issue may be solved through source separation on the ambient audio stream from the integrated headphone microphones. Sirens as an example for critical sound sources may be isolated from the auditory scene and played back on the headphones. This falls into the category of Selective

Hearing (SH) which has seen many advancements in terms of detection, classification, separation, localization and enhancement of sound sources [1].

Existing DSP-based source separation approaches that may be used for SH applications are commonly based on statistical means. The common goal is the estimation of the inverse of the mixing matrix  $A$  which was used to mix  $N$  real source vectors  $s(t) = (s_1(t), \dots, s_N(t))^T$  to  $M$  output vectors  $x(t) = (x_1(t), \dots, x_M(t))^T$ :

$$x(t) = A s(t) . \quad (1)$$

Since both  $A$  and  $s(t)$  are unknown, finding  $A^{-1}$  is an ill-posed problem. Furthermore, in the context of practical applications  $M \ll N$ . The solution to this problem is therefore approximated under various assumptions [2].

Independent component analysis is a statistical source separation method under the assumption that the sources are



statistically independent and identically, but non-Gaussian distributed [2]. A further common separation approach is non-negative matrix factorization which assumes that the mixing matrix  $A$  and the sources  $s$  are non-negative [2]. Regarding more recent approaches, projection-based demixing was introduced [4]. In this approach, the observed output mixture  $x(t)$  is decoded into tensors of time, frequency and channels. Then, different spatial projections are created within these tensors to identify individual sources.

In contrast to the previous methods for DSP-based source separation, the use of a neural network architecture poses a more promising application-focused approach for the separation of sirens in particular. Since sirens are usually simple combinations of sinusoidal sounds, neural networks can be trained to detect and separate them from a diffuse sound ambience. This poses no further constraints on the audio material except a certain level of presence of the siren, which may be negligible within everyday boundaries.

In this context, we compare a CNN and LSTM model embedded into a corresponding system architecture for the extraction of sirens from a stereo time signal, without digital signal processing requirements or assumptions. The data preparation step simulates signals as they may arrive at microphones integrated into headphones. Apart from accuracy metrics, the proposed system is tested in a simulated traffic scenario. The acoustic scene is generated via the object-based audio approach of the Fraunhofer SpatialSound Wave technology [5].

## 2. Data Collection

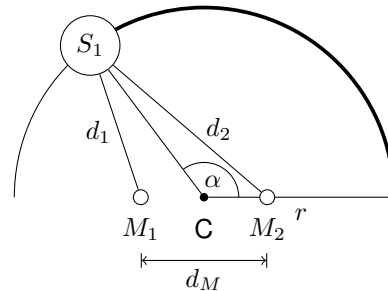
A diverse set of training and validation data was collected for training and evaluation of the DNN models. About 200 Gigabytes of online audio material were scanned for the curation of two final datasets of 20 Gigabytes in total size. One dataset was used for training, the other one for testing. In addition to the evaluated online content, free field recordings were made to further expand the dataset. The audio material was sampled at 44.1 kHz and a resolution of 24 bit.

The training and test dataset are comprised of two parts: ambience sounds and siren sounds. The curated ambience sounds mostly consist of various traffic ambience recordings, as these represent the setting where siren sounds most likely occur. Apart from traffic recordings, other typical city and outdoor ambiences were included, such as from parks, malls, train stations and similar public places. A minor part of the curated ambience sounds were obtained from the UrbanSound data set [14]. In order to challenge the DNN models, white noise and pink noise sequences, as well as ambience recordings with sounds present in a frequency range similar to sirens were added, e.g. twittering birds and playing children.

The majority of the curated siren sounds are free from strong artificial or recorded reverb as well as delay effects to ensure network model training without data pollution. For further data cleansing, the siren sounds were high and low cut at 150 Hz and 7500 Hz, respectively, in order to remove low frequency rumble and irrelevant high pitched noise.

## 3. Spatial Data Simulation

Signals as they may arrive at microphones placed on the outside of ear cups were simulated using a custom acoustic simulation script. This allowed for automated time delay and distance dependent gain calculations using two virtual microphones, as shown in Figure 1.



**Fig. 1:** Virtual microphone simulation. The position of the sound source  $S_1$  is defined by the radius  $r$  and the azimuth  $\alpha$ . Depending on the distances  $d_1$  and  $d_2$  between  $S_1$  and the microphones  $M_1$  and  $M_2$ , the gain and time delay for each microphone signal is calculated.

The simulation works as follows. Two virtual microphones  $M_1$  and  $M_2$  are placed at a distance  $d_M$  from one another, centered around  $C$ . The position of a virtual sound source of interest  $S_1$  is defined by the radius  $r$  and the azimuth  $\alpha$ , alongside its initial gain  $g(S_1)$ . The gain of the  $S_1$  signal arriving at  $M_1$  and  $M_2$  is then calculated using the distances  $d_1$  and  $d_2$  between the microphones and  $S_1$  via equation 2:

$$g(S_1, M_{1|2}) = \frac{g(S_1)}{d_{1|2}^2}. \quad (2)$$

Additionally, the time delay  $\Delta t$  for the signal from  $S_1$  arriving at each microphone is calculated using equation 3 with the speed of sound  $v_s$  at 343 m/s:

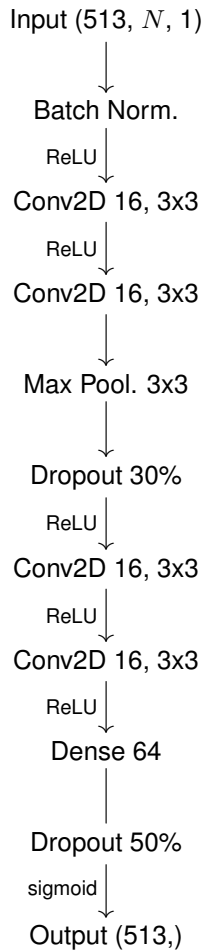
$$\Delta t(S_1, M_{1|2}) = \frac{d_{1|2}}{v_s}. \quad (3)$$

The stereo mix of the mono microphone signals creates the impression of the sound source of interest coming from the intended position. The distance  $d_M$  can be adapted to the outer distance between the ear cups of different headphones. The simulation may be improved in the future through the simulation of the human head, since with this setup a distance  $d_M \geq 1$  m led to the most realistic simulations [12].

## 4. Neural Network Models

A CNN and LSTM model were evaluated for the task of reliable separation of sirens. The models were implemented using the Keras API of Google Tensorflow [3] and embedded into the corresponding system architecture presented in section 5.

The proposed CNN model is based on the model introduced in [10]. It is fed with 25 chronological STFT windows in order to predict an STFT mask for the central 13th STFT window which is then used to separate the source of interest.



**Fig. 2:** Spectrum masking model. The model is fed with packages of  $N$  STFT windows. After batch normalization, one output STFT window mask is predicted through multiple convolutional, dropout, dense, and one max pooling layer.

The original model from [10] was optimized in several steps, resulting in the final model shown in Figure 2. In comparison to the model in [10], the leaky ReLU activation functions were substituted by regular ReLU activations [16]. The dropout rates were increased to 30% to counteract model overfitting [15]. Apart from increased dropout rates, a batch normalization layer was added to normalize the input before the convolutional layers [7]. This yielded better model predictions on a larger variety of input data during early testing. From a practical standpoint, this also produces a more constant output volume of the separated sound source.

In a last step, the stochastic gradient descent optimizer of the original model with custom parameters was replaced by the Adam optimizer [9]. This decreased training duration and further improved model performance. Due to exploding gradients, a clip value of 3.0 was set for the optimizer [13].

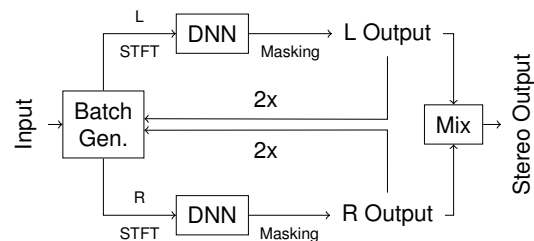
The input of the network as denoted in Figure 2 has the two-dimensional shape  $513 \times N \times 1$ , where the 513 frequency bins result from the STFT window size of 1024 samples, as further explained in section 5. The  $N$  windows correspond to the 25 chronological STFT windows of the original model from [10]. This time context of  $N = 25$  was changed to

$N = 17$  windows in order to evaluate network predictions with less data. These results are presented in section 6.1.

For potential performance improvement, the chronological context of each predicted STFT window was also replicated by an LSTM model [6]. The tested LSTM model uses all layers of the CNN model in Figure 2 before the output layer as time distributed input to an LSTM layer with 64 LSTM cells. The metrics of this alternative model are also discussed in section 6.1. Despite significantly longer training times, no practical prediction improvements compared to the CNN model were identified.

### 5. System Architecture

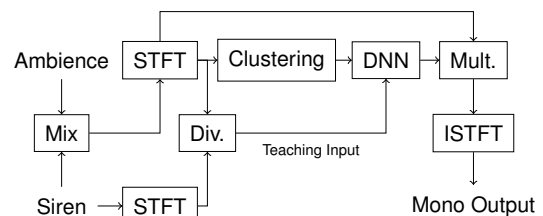
The complete system architecture for the separation of sirens is depicted in Figure 3. One instance of the DNN is applied



**Fig. 3:** Stereo separation pipeline. The stereo ambience mix is split into left and right channels by the batch generator. The DNN for separation is then applied three times to each channel individually. The mono output of the two network instances is mixed back together to obtain the final stereo output.

to each channel of the stereo input signal. The correct localization of the siren signal is preserved through the final mix of the two separated mono signals.

As presented in the previous section, the CNN and LSTM models operate on data in the frequency domain in order to learn spectrum masks. These masks are applied to the STFT windows of the mixed ambience input, which separates the siren by isolating corresponding frequency bins. The final processing pipeline for this approach is depicted in Figure 4. The network is trained with a processed ambience mix, and



**Fig. 4:** Spectrum masking pipeline. The ambience mix and siren signal are converted to the frequency domain. The DNN is trained to predict spectrum masks to separate the siren signal from the mono ambience mix.

a processed version of the siren signal as the teaching input. Both the mix and the siren signal are first transformed into the frequency domain via STFT. The transformation uses the Hann window function on a window size of 1024 samples and

a window overlap of 512 samples. This results in complex STFT windows in the shape  $1024 \times 1$ .

Since the neural network only operates on real data, the magnitude spectrum of each STFT windows is used for further processing. This results in spectrum windows in the shape  $513 \times 1$ . For the teaching input of the network, the magnitude windows of the ambience mix and the siren signal are divided by one another to obtain the desired spectrum masks. The calculation of each spectrum mask  $\vec{m}$  for the respective ambience and siren STFT windows  $\vec{w}_a$  and  $\vec{w}_s$  is denoted in equation 4:

$$\vec{m} = \begin{cases} \frac{|\vec{w}_s, i|}{|\vec{w}_a, i|} & \text{if } \vec{w}_a, i \neq 0 \\ 0 & \text{else} \end{cases} \quad \text{for } i = 1, 2, \dots, 513. \quad (4)$$

Instead of training the network to predict one spectrum mask for one spectrum window as its input, a clustering step provides the network with more time context. The clustering step adds  $\frac{N-1}{2}$  windows before and after the STFT window of interest to the network input. The network thus receives packages of magnitude windows in the shape  $513 \times N \times 1$ . It then outputs a mask in the shape  $513 \times 1$  for the  $\frac{N+1}{2}$ th input window, with all mask entries lying in the interval  $[0, 1]$ .

Since the network is only fed with the magnitudes of the complex STFT windows, the original phase information of the complex STFT windows is added back to the spectrum windows before the spectral masks are applied. The separated time signal can then be calculated using the inverse short-term Fourier transform.

## 6. Evaluation

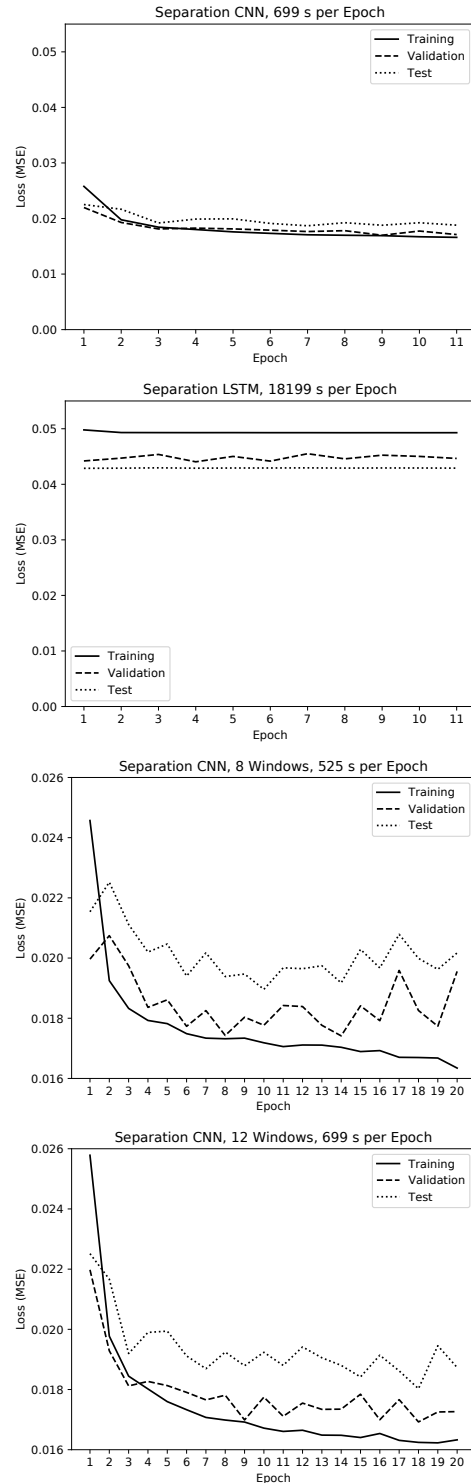
The presented DNN models were evaluated in a quantitative and qualitative manner. The different metrics as well as the separation results from a realistic acoustic scene simulation are presented in the following.

### 6.1. Network Model Metrics

The final training metrics of the CNN and LSTM models are shown in Figure 5. It is apparent that the CNN with a time context of eight STFT windows converges to consistently lower loss metrics in terms of training, validation and testing compared to the LSTM model. Taking into account the significantly longer training duration of a median of 18199 seconds for the LSTM model compared to the 699 seconds for the CNN model, the CNN clearly achieves better performance.

Since alerts, including sirens, are typically short sounds that stand out from an auditory scene, it was tested if reducing the number of STFT windows fed into the CNN could decrease training and prediction durations without compromising model accuracy. Figure 5 shows training results with eight STFT windows compared to twelve STFT windows.

The CNN with a time context of eight STFT windows shows clear signs of overfitting after about 13 epochs. Conversely, the model with eight windows seems to converge slightly sooner and trains about 25% faster with respect to the median training times of 525 seconds and 699 seconds, respectively.



**Fig. 5:** Training results. The first two charts show the less accurate results of the LSTM model despite a roughly 25 times longer training time compared to the CNN model. The third chart shows the CNN loss trend with a time context of eight STFT windows, the bottom chart for a time context of twelve STFT windows. For each scenario, the median training time for one epoch is denoted above.

After 18 epochs, however, the CNN with a time context of twelve windows reaches a loss minimum below all loss minima of the eight window time context model. The CNN was therefore identified as the better performing model.

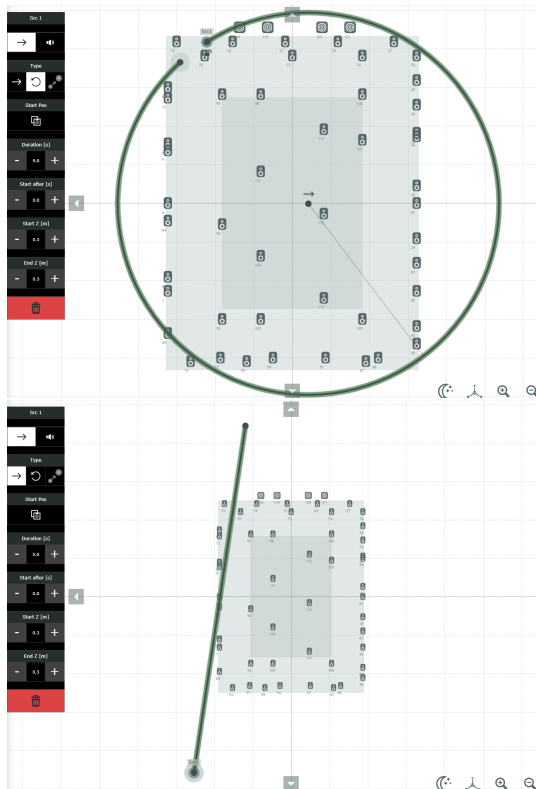
## 6.2. Spatial Audio Scene Simulation

For a realistic evaluation of the system with the better performing CNN model, recordings in a large test room with a three-dimensional speaker arrangement were made. The room dimensions are 9 m × 7.1 m × 4.7 m (L×B×H). The speaker setup consists of 49 satellite speakers and four subwoofers.

The Fraunhofer SpatialSound Wave technology was used to play back three-dimensional ambience recordings on the speaker setup via the object-based audio approach [5]. Besides the realistic playback of the ambience recordings, the setup enables the mapping of a siren signal to an audio object which can arbitrarily be moved in space.

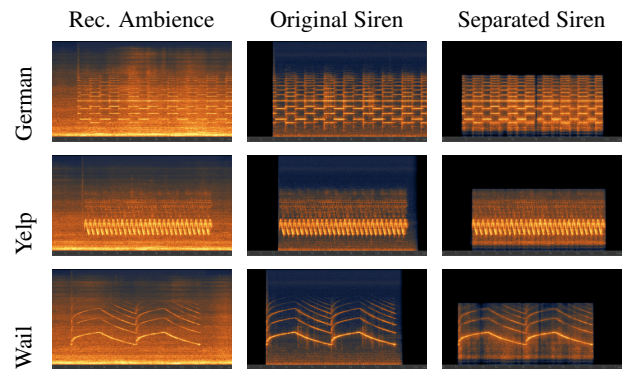
Two condenser microphones were placed at the acoustic hot spot of the demo room. A distance of 0.2 meters between the microphones was chosen for a distance similar to small microphones as they could be attached to the back of ear cups.

Three different sirens were each moved along a circular and linear audio object path as shown in Figure 6. The

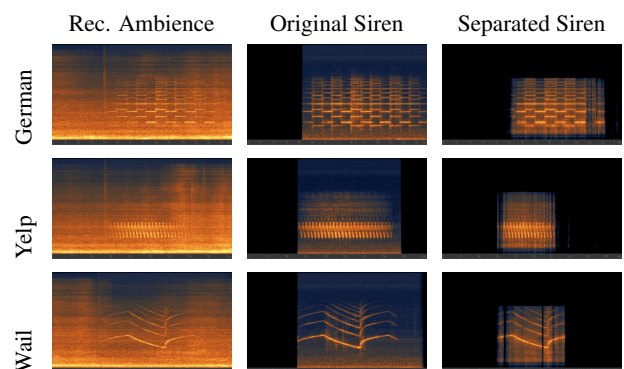


**Fig. 6:** Automated siren movements. The sirens are moved along the 360° circular path at the top to explicitly test localization accuracy. The linear path at the bottom simulates sirens passing by in traffic.

sirens represent a German, American Wail and American Yelp police siren. Common traffic noise was played back from a three-dimensional recording to simulate a realistic traffic ambience. The volume ratio between the sirens and traffic ambience was adjusted by ear. For the circular siren movement, a constant siren gain was chosen to explicitly test the localization accuracy of the CNN model at every azimuth  $\alpha$ . In the case of the linear siren movement, the volume of the siren was automated to be distant dependent, i.e.



**Fig. 7:** Separation results for circular siren movements. The CNN produces reliable results apart from short interruptions in the German and wail siren signals.



**Fig. 8:** Separation results for linear siren movements. The louder parts of the sirens are separated successfully, while the quieter parts are not always recognized.

increasing towards the microphone position and decreasing while moving away from it.

Separation results for the circular siren movement are shown in Figure 7. The spectrograms are plotted on the Mel frequency scale using the monophonic sum of the signals [11]. The separated sirens are close to the original siren signals and the sections without sirens are successfully kept quiet by the network. The slightly brighter areas around the main sinusoidal siren sound waves indicate a low remaining noise floor in the separated sequences.

Separation results for the linear siren movement path are plotted in Figure 8. Contrary to the separation results of the constant siren volume on the circular movement path, these results reveal unreliable separations for the quieter siren sections. While the German siren is recognized throughout most of its occurrence in the recorded ambience mix, both the yelp and wail siren are only separated properly during the louder sections, i.e. when the virtual distance of the siren decreases towards the microphone position. This indicates that the CNN is not able to reliably detect sirens below a certain volume threshold.

Although quieter siren signals appear to remain a challenge for the tested models, quiet sirens may also be less of importance if they are far away. Further simulated or real-life

tests will therefore be needed to reliably identify the detection boundaries of the proposed architecture and optimize the presented system and training data accordingly.

## 7. Conclusion

This paper compared a CNN and LSTM model with a corresponding system architecture for the separation of sirens. The CNN model training is quicker and produces more accurate results. The CNN model was also evaluated from a practical standpoint using the Fraunhofer Spatial SoundWave technology for an object-based traffic scene playback. The circular movement of three typical sirens around the microphone position showed reliable separation results. The linear movement simulation revealed detection boundaries for siren signals below a certain volume threshold.

Future research may reveal more efficient architectures that meet the limited system resources and realtime requirements of an embedded system inside noise-cancelling headphones. Processing delays will need to be kept within strict time constraints, for instance in dangerous traffic situations, while conserving system resources, e.g. with respect to battery life. For a guaranteed reliable performance in difficult situations, the exact boundaries of the proposed system will need to be identified and corresponding training data optimizations will need to be made.

## 8. References

- [1] Estefania Cano and Hanna Lukashevich. 2019. Selective Hearing: A Machine Listening Perspective. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE.
- [2] Pierre Comon and Christian Jutten. 2010. *Handbook of Blind Source Separation: Independent component analysis and applications*. Academic press.
- [3] Mark Daoust. 2019. Keras. <https://www.tensorflow.org/guide/keras>. [Online; accessed 21-August-2019].
- [4] Derry Fitzgerald, Antoine Liutkus, and Roland Badeau. 2016. Projection-based demixing of spatial audio. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 24, 9 (2016), 1556–1568.
- [5] Alejandro Gasull Ruiz, Christoph Sladeczek, and Thomas Sporer. 2015. A description of an object-based audio workflow for media productions. In *Audio Engineering Society Conference: 57th International Conference: The Future of Audio Entertainment Technology—Cinema, Television and the Internet*. Audio Engineering Society.
- [6] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [7] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).
- [8] Tominori Kimura. 2011. Noise-cancelling headphone. US Patent 8,045,726.
- [9] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [10] Alejandro Koretzky, Karthiek Reddy Bokka, and Naveen Sasalu Rajashekhara. 2018. Real-time audio source separation using deep neural networks. US Patent 10,014,002.
- [11] Beth Logan et al. 2000. Mel Frequency Cepstral Coefficients for Music Modeling.. In *ISMIR*, Vol. 270. ISMIR, 1–11.
- [12] John C Middlebrooks and David M Green. 1991. Sound localization by human listeners. *Annual review of psychology* 42, 1 (1991), 135–159.
- [13] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2012. Understanding the exploding gradient problem. *CoRR, abs/1211.5063* 2 (2012).
- [14] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 1041–1044.
- [15] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.
- [16] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. 2015. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853* (2015).



## Full Reviewed Paper at ICSA 2019

Presented by VDT.

### Exploring Audiovisual Support Systems for In-Car Multiparty Conferencing

F. Weidner<sup>1</sup>, B. Fiedler<sup>1,2</sup>, J. Redlich<sup>1,2,3</sup>, and W. Broll<sup>1</sup>

<sup>1</sup> *TU Ilmenau, 98693 Ilmenau, Germany, Email: florian.weidner@tu-ilmenau.de*

<sup>2</sup> *Fraunhofer IDMT, 98693 Ilmenau, Germany, Email: bernhard.fiedler@idmt.fraunhofer.de*

<sup>3</sup> *ITA Weimar mbH, 99428 Weimar, Germany, Email: ita@ita-weimar.de*

#### Abstract

Calling while driving poses a severe safety risk. When more than two people are involved in a call - a conference call - this risk increases even more. Intelligent vehicles could offer support systems that ease the cognitive burden of such a multiparty calls. We explore the possibilities of such advanced driving assistant systems (ADAS) in two ways: first, we investigate object-based spatial audio where each remote caller is modeled as a distinct audio source. Second, we apply a non-intrusive ambient stereoscopic 3D (S3D) visualization that indicates the current speaker and its location. In a between-subject design driving simulator study (n=56), we assess workload, user experience and driving performance. Surprisingly, we found no positive effect of object-based audio. However, we present evidence how a supporting visualization might lower situational stress and increase the system's dependability. We conclude that a supportive and intelligent stereoscopic visualization is a promising candidate for enhancing multiparty conference calls while driving.

## 1. Introduction

People work in their cars while driving. They search through papers, make notes, check emails, or schedule meetings [18, 25]. It is likely and anticipated that advances in connectivity and automation will increase the time people work in cars and will make other tasks more likely to be performed [6, 21]. In addition to the mentioned work-related activities, people make business calls while driving [18, 25]. They call clients, secretaries, or colleagues to productively use the time they spent in their car. However, having a phone call while driving is, despite all technological advances and hands-free technology, a dangerous task. Leung et al. [19] concluded that making a cognitively demanding hands-free phone call while driving is as risky as driving with a blood alcohol concentration of 0.07 to 0.10%. Moreover, as soon as phone calls have multiple remote participants (multiparty conversation), cognitive load increases [29]. In particular, it gets harder to distinguish the speakers' voices and by that, to

follow the conversation [8]. In a face-to-face meeting, it is easy to distinguish who is speaking: we see the people and hear their voices from where they are sitting or standing. In a remote multiparty conversation however, the sound sources that represent the communication partners usually come from the same direction. Additionally, there is no visible representation of the callers. Restoring both properties - spatial sound and visual representation - might help to lower the cognitive burden of such remote multiparty conversations and make them less risky.

Thus, this paper investigates two potential solutions to enhance the experience during multiparty conference calls while driving: an object-based spatial sound system and an adaptive visualization. The former enables the listener to locate each voice at the correct angle in the room. The latter provides a non-intrusive ambient stereoscopic 3D (S3D) dashboard visualization, adding visual representations to the voices, and enables the user to match the voice to a name.



The rest of the paper is structured as follows: We start with background information and related work (Section 2.1). Following, we present our experimental design and prototype in Section 3. Section 4 presents our results and Section 5 interprets and discusses them. Finally, we conclude and provide a brief outlook in Section 6.

## 2. Background & Related Work

### 2.1. Stereoscopic 3D

Traditional displays only show one image for the left and right eye. Depth is visible, but only via cues like occlusion or linear perspective. Stereoscopic 3D (S3D) visualizations display individual images for the left and right eye. The human visual system then fuses these images and calculates binocular depth cues for depth perception similar to real-life [9].

### 2.2. Channel- & Object-based Audio

Audio scenes usually consist of many channels mixed together by the audio engineer. For channel-based audio (CBA), the audio scene is down-mixed regarding standardized speaker layouts [8]. Since speaker layouts between common reproduction systems and the audio engineer's production or recording system usually do not match, the originally mixed sound field can not be reproduced exactly. This leads to a loss of spatial information (e.g. all sound sources come from one direction).

In contrast, object-based audio (OBA) does not store pre-produced audio channels, but instead sound objects consisting of audio and metadata. Knowing the positions of loudspeakers, the sound field of the audio scene can be reproduced exactly by calculating each speaker signal independently in real time during playback [2, 22]. Hence, spatial information is retained (e.g. the direction of speakers during a multiparty conversation). Similar results can be achieved by applying a blind source separation algorithm that decomposes a multi-channel input stream into several output streams [23]. Although channel-based and object-based audio are both spatial audio formats, only the object-based approach enables the dynamic, real-time capable spatial location of sound sources. Therefore in this work only the object-based approach reproduces spatial audio.

### 2.3. Multiparty Conference Calls

Many research groups try to lessen the cognitive demand of calling while driving by exploring novel technologies like AR-glasses [15, 16] and using video-call software [14]. However, these systems have only been tested during one-to-one conversations.

For multiparty conference calls, Rajan et al. [26] try to solve this issue by using an intelligent user interface. It successfully performs, among other measures, speaker identification and adds presence indicators in form of personalized background noises to the conference call in a desktop setting. Kilgore reports a lower level of perceived difficulty when using spatial audio in a desktop audio conference system [13]. Their system increases the quality of the conference call but was not applied in the more safety-critical automotive context. No significant benefit of spatial audio was found by Inkpen

**Tab. 1:** Experimental design with four groups.

	No visualization	Stereoscopic 3D
Channel-based audio	CBA-NoV	CBA-S3D
Object-based audio	OBA-NoV	OBA-S3D

et al. [10]. However, they found that displaying a visual indicator in form of spatial video improved people's ability to follow the conversation. It is important to note that the application of visual indicators during manual driving has to be handled carefully because the main focus of the driver should be on the traffic and surroundings. Wickens' theory of shared resources argues against providing visual cues while driving because driving is a highly visual task and the resources for processing visual information should be allocated for the driving and not for a secondary task [32]. However, recent advances in ambient lighting (e.g. Loecken et al. [20]) provide an interesting possibility to implement a non-intrusive vision-based support system for multiparty conference systems. Also, previous work on stereoscopic 3D dashboard visualizations showed that they do not necessarily decrease driving performance if designed carefully - especially for change detection tasks [30, 31]. Furthermore, S3D can improve user experience while driving when designed carefully [3, 4].

The high mental workload induced by phone calls motivated us to explore spatial audio and stereoscopic 3D visualizations as support systems during such calls more closely.

## 3. Study

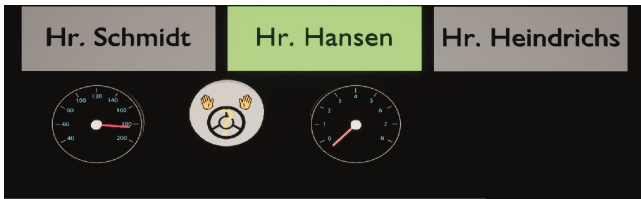
Our study focuses on multiparty conference calls. The situation resembles a remote job interview where several employees of a company talk with a potential new employee who is currently driving on a highway.

### 3.1. Experimental Design

The experiment had a between-subject design. Dependent variable is the type of user interface. Participants experienced one of the four user interfaces as indicated in Table 1. Independent variables are described in Section 3.4.

In *CBA-NoV*, the three speakers' voices come from the same direction in front of the user and no supporting visualization was offered. This condition acts as the baseline. In *OBA-S3D*, the three speakers' voices can be perceived from distinct directions (left, center, and to the right of the driver) while supporting S3D visualization was offered.

We chose a between-subject design for two reasons: first, to shorten experiment duration. Second, a mixed design would have required a second simulated conversation. It turned out to be a challenging task to design two similar conversations that induce the same workload and therefore allow a comparison of all conditions, but do not repeat themselves. Hence, we applied the between-subject design that requires more participants but allows easy comparisons across groups.



**Fig. 1:** The user interface adapts to the number of participants. The speaking person is highlighted using a green color. Location of tags matches position of speakers in auditory scene for OBA-conditions.



**Fig. 2:** Default view of the driving simulation.

### 3.2. Sample

The final sample consisted of  $N = 56$  participants (male = 38 or 67.9%, female = 18 or 32.2%) with a mean age of 30.5 years ( $SD = 8.22$ ). All possessed a valid driving license, had no hearing impairments, had normal or corrected-to-normal vision, and passed a stereopsis test if they were assigned to a S3D group (Random Dot Stereogram, [28]). 45 out of 56 (80.36%) had experience with stereoscopic displays and 26 (46.43%) had experience with 3D Audio. Overall, 31 participants (55.36%) said that they regularly make phone calls while driving. Mean MSSQ score is 9.53 ( $SD = 8.62$ ).

### 3.3. Apparatus

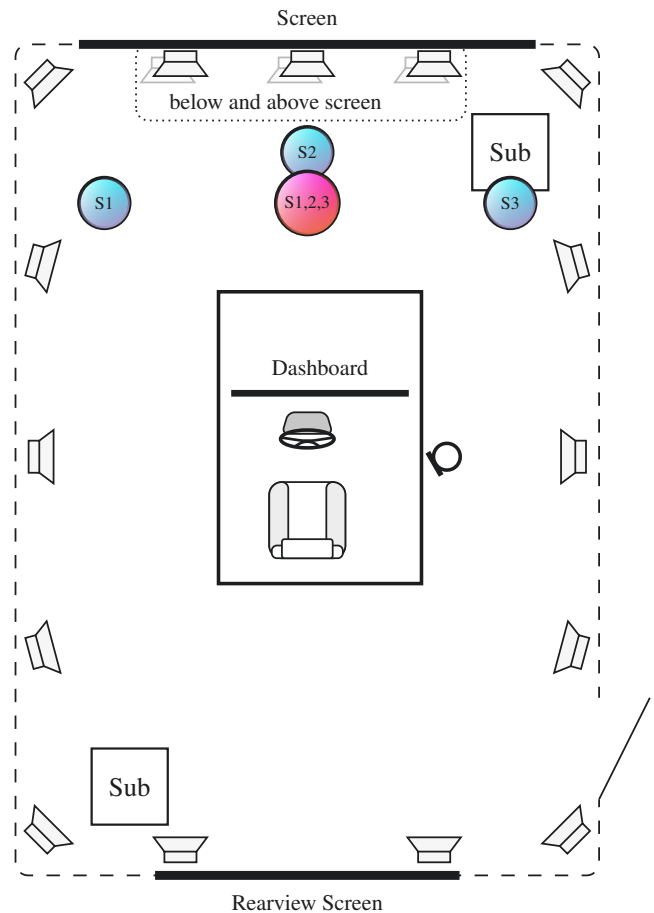
Figures 2 and 3 illustrate the simulation environment. A screen (2D, 3.6×2.26 m, 2560×1600 pixel at 120 Hz) displays the driving environment. The car mock-up has an integrated spatially augmented reality dashboard (L-shaped, 90×60 cm, 2560×1600 pixel at 120 Hz, in S3D mode: 60 Hz per eye). It is further equipped with a Thrustmaster TX Racing Wheel Leather Edition and pedals. Volfoni Active 3D glasses<sup>1</sup> enable the stereoscopic visualization. An Optitrack tracking system<sup>2</sup> realizes head tracking and the head-coupled perspective. The rear-view was visible via a real car mirror and a 1920×1080, 30 Hz projection. All visualizations were realized using Unreal Engine 4.18.3<sup>3</sup>. Field of view was set to 70°.

Figure 1 illustrates the user interface supporting the multiparty conversation. The name tags appear one by one according to the callers in the multiparty conversation. The name tags and colors are carefully calibrated to be non-intrusive and non-glaring while driving. They are located at the top of the

<sup>1</sup><http://volfoni.com/>, 2019-01-03

<sup>2</sup><http://optitrack.com/>, 2019-01-03

<sup>3</sup><http://unrealengine.com/>, 2019-01-02



**Fig. 3:** Overview of simulation environment. Positions of dialogue partners marked for OBA in blue and CBA in red.

dashboard so that they can act as an ambient lighting cue. Disparity of the name tags is  $D = 0,218^\circ$ . That means that they appear 4 cm behind the projection plane with a viewing distance of 80 cm and an interpupillary distance of  $IPD = 63$  mm. Disparity of the speedometer, tachometer, and cruise control indicator is  $D = 0^\circ$ . Stereoscopic 3D was chosen as an additional cue for the ambient visualization.

Locations of the name tags correspond to the intended seat distribution of the dialogue partners. In the OBA-condition, speakers' voices were located at the intended position relative to the driver. In the CBA-conditions, all three voices come from the same direction directly in front of the user resembling mono audio playback.

To reproduce audio, the wave field synthesis based 3D audio system SpatialSound Wave<sup>4</sup> (SSW) using 18 Seeburg TS-nano speakers and two Seeburg TSM subwoofers was installed. The speakers are positioned in a horizontal plane around the listening position at a height of 1.4 m. See Figure 3 for speaker positions. The front speakers are positioned above at 2.5 m and under the screen at ground level not blocking view to the screen. Using delay and gain adjustments, the sound sources can still be perceived in the horizontal plane.

<sup>4</sup><https://www.idmt.fraunhofer.de/en/institute/projects-products/spatialsound-wave.html>, 2019-01-30

The audio system was calibrated and equalized to compensate room acoustics and speakers' characteristics.

Synchronization between visualization using Unreal Engine and auralization using SSW was implemented by exchanging scene data graphs as XML via UDP. By that, all sound sources (environment, other cars, engine and tires of ego car, wind noise and voices of the speakers) are managed in real-time and positioned in 2D space around the driver. Audio reproduction also included doppler effect of passing objects like cars.

### 3.4. Measures

To describe the sample, we apply the motion sickness susceptibility questionnaire (MSSQ, Golding et al. [7]). We further ask for simulator sickness using the Simulator Sickness questionnaire (SSQ, Kennedy et al. [12]) and for user experience using the user experience questionnaire (UEQ, Laugwitz et al. [17]). In order to assess workload, we apply the Driver Activity Load Index (DALI, Pauzie et al. [24]). The DALI is a modified version of the NASA TLX [?], especially designed to assess workload of drivers. Driving performance was operationalized using number of steering reversals and number of lane departures (measured using widest part of the vehicle) [27]. We count a steering wheel reversal when the driver turns the steering wheel 6 degrees in one direction and within 2 seconds 6 degrees in the opposite direction. We count a lane departure when the car crosses the center of a lane marking without making a lane change.

### 3.5. Procedure

When participants arrived, they filled out a consent form, the MSSQ, and a general questionnaire. They were randomly assigned to one of the four groups. They took a seat in the driving simulator and received general information about the phone interview, the car, and its capabilities. Depending on their assigned group, participants experienced different audiovisual user interfaces (c.f. Section 3.1). However, all wore stereoscopic 3D glasses. The virtual car was equipped with cruise control and participants were instructed how to use it. They were further told to respect traffic rules. In our study, participants had to drive along a 12 km curved three-lane highway with low traffic (approximately 5 cars per km; primary task) and engage in a multiparty-conversation (secondary task). After about 3 km, which were considered training, the phone rang. When participants had accepted the call with a button on the steering wheel, a simulated job interview started. In our scenario, the participant had previously applied for an internship at the fictional "Institute for Thermodynamics". Three virtual members of the institute took part in the interview. For playing questions, answers, and other sounds (e.g. agreeing sounds) of the callers, we used a wizard-of-oz-based audio player for multiparty conversations<sup>5</sup> which was controlled by an operator from another room. Approximately 10 seconds after they had hang up, the drive ended. After that, they filled out the UEQ, SSQ, and DALI. That concluded the experiment. Total experiment duration was approximately 25 minutes. Participants were not paid, but had the chance to win 50 Euro.

<sup>5</sup><https://github.com/JoReIMT/Dialogue-Sample-Player>, Dialogue-Sample-Player, GPL-3.0, 2019-03-01

## 4. Results

All data follows normal distribution (tested with Shapiro-Wilk test, histograms, and QQ-plots, [33]) and shows homogeneity (tested with Levene's test) if not stated otherwise. An  $\alpha$ -value of 0.05 was used as significance criterion when necessary. Data was analyzed using R 3.5.2 (afex 0.22.1, fBasics 3042.89, bestNormalize v1.3.0, and MASS 7.3-51.1).

The overall drive lasted about 6 minutes and 4.97 seconds ( $SD = 75.07$  seconds;), depending on the answers given by the participants. Mean driving speed was  $M = 116.47$  km/h ( $SD = 8.00$  km/h). Participants drove on average 11.81 km.

### 4.1. Simulator Sickness

Data of the SSQ is not normal distributed. Hence, we applied a Yeo-Johnson transformation to correct for normality [34]. A two-way ANOVA found no evidence for significant differences in simulator sickness on Yeo-Johnson transformed data of the SSQ (untransformed data:  $M_{Nausea} = 18.91$ ;  $SD = 17.59$ ;  $M_{Oculomotor} = 18.27$ ;  $SD = 15.55$ ;  $M_{Disorientation} = 14.42$ ;  $SD = 21.21$ ;  $M_{Total} = 20.30$ ;  $SD = 17.51$ ;  $p > .340$ ).

### 4.2. User Experience

Results of the UEQ are presented in Figure 4. Results of a two-way ANOVA suggest that there is a significant interaction effect on the *Attractiveness* scale ( $F(1, 52) = 6.82$ ,  $p = .012$ ,  $\eta_p^2 = .12$ ). Closer inspection using Tukey's HSD confirms that CBA-S3D ( $M = 0.88$   $SE = 0.09$ ) was perceived significantly more attractive than CBA-NoV ( $M = 0.46$   $SE = 0.08$ ;  $t(52) = 2.858$ ,  $p = .0302$ , Cohen's  $d = 1.26$ ) as Figure 5 indicates. There is also a significant main effect of the video condition on *Dependability* stating that the S3D condition was perceived more dependable than the NoV condition (S3D:  $M = 0.60$ ,  $SE = 0.12$ ; NoV:  $M = 0.24$ ,  $SE = 0.12$ ;  $F(1, 52) = 4.31$ ,  $p = .043$ ,  $\eta_p^2 = .08$ ). No other main or interaction effects were found.

### 4.3. Driver Activity Load Index

Results of the DALI questionnaire are shown in Figure 6. Data is not normal distributed and data transformations do not lead to a reasonable normal distribution. Hence, we decided to analyze data using multiple Wilcoxon Rank Sum tests. Because of the large standard deviations and rather small sample size, we deliberately did so without correction for multiple comparisons. This allows for data exploration, detection of possible effects, and making suggestions for future research. However, interpretation of any significant differences requires taking into account this very liberal procedure.

On the *Visual Demand* scale, there is a significant difference between OBA-NoV and OBA-S3D ( $W = 52$ ,  $p = .035$ ,  $r = .281$ ). indicating that OBA-NoV ( $Mdn = 48.5$ ) was less visually demanding than OBA-S3D ( $Mdn = 15.0$ ). On the *Situational Stress* scale, we found evidence for a significant main effect of visualization between S3D and NoV with  $W = 266.5$ ,  $p = .394$ ,  $r = .276$  suggesting that the S3D condition ( $Mdn = 19.0$ ) leads to less stress than the NoV condition ( $Mdn = 39.5$ ). For all other comparisons, test results are non-significant with  $p > .079$ . This tells us that the other sub-scales measured by the DALI do not significantly differ between groups and conditions.

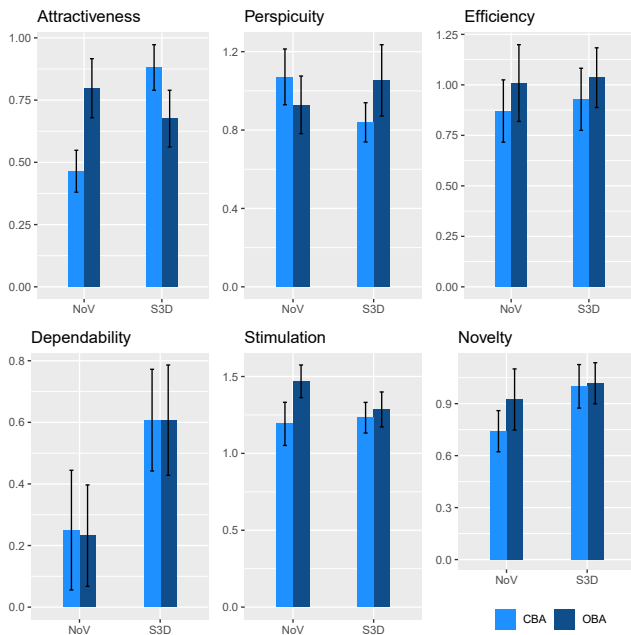


Fig. 4: Means and standard errors for the UEQ ([-3;3], higher is better).

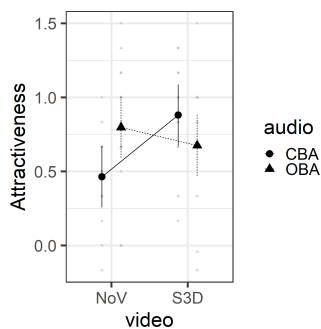


Fig. 5: Interaction plot for “Attractiveness” of the UEQ. CBA-S3D was perceived more attractive than CBA-NoV

Because some authors suggest that the ANOVA is very robust to violations of the normality assumptions (e.g. Field et al. [5]), we ran a two-way ANOVA for the two found effects. The more conservative ANOVA confirms the main effect of the visualization on *Stress* ( $F(1, 52) = 4.75, p = .034, \eta_p^2 = .08$ ) but not the interaction effect on *Visual Demand* ( $F(1, 52) = 3.30, p = .075, \eta_p^2 = .06$ ).

#### 4.4. Driving Performance

##### 4.4.1. Number of Steering Wheel Reversals

We applied a Tukey transformation to establish normal distribution. A two-way ANOVA found no significant effects in steering wheel reversals ( $p > .406$ ). By that, we can assume that the different audiovisual support systems did not significantly influence this measure of driving performance.

##### 4.4.2. Number of Lane Departures

Data on number of lane departures does not follow a normal distribution so we applied a Tukey transformation. Again, a two-way ANOVA found no significant effects in the Tukey-transformed number of steering wheel reversals ( $p > .582$ ). This indicates that there is insufficient evidence for an influ-

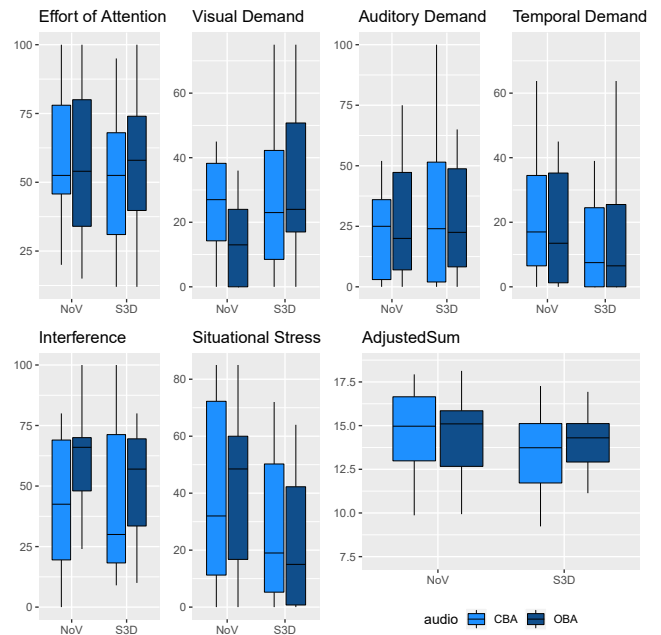


Fig. 6: Medians for the scales of the DALI ([0;100], lower is better).

ence of the support systems on the number of steering wheel reversals.

## 5. Discussion

We explored two support systems: an object-based audio which emits the speakers’ voices from distinct directions and a stereoscopic 3D dashboard visualization providing indications who is speaking at the moment via non-intrusive ambient lighting and name tags. We expected that each support system on its own leads to less workload and higher user experience (UX). We further expected that the combination of both support systems outperforms each support system regarding workload and UX.

Results indicate that participants’ responses to the different support systems are very conservative. We did not find a clear favorite with respect to user experience or workload.

Prior research indicates that spatial sound makes it easier for people to follow a conversation and reduces perceived difficulty in multiparty conversations [1]. In our study, spatialized audio reproduction did not significantly improve driving performance or reduce the drivers’ workload during phone conference. Whereas prior research on multiparty conversation was done without any ambient noise [1, 29], the complete auditory scene including environmental sounds was reproduced with spatial audio in our study. Maybe the differences between spatialized speech and spatialized noise were not prominent enough to affect the measures positively.

Nevertheless, results indicate that perceived *Situational Stress* is lowered by the S3D support system. As mentioned in Section 4, this is based on a liberal test procedure. However, it acts as a strong motivation for further research in this area.

SAE International mentions values for comparison of lane departures as guidance ranging from 7.1 to 16.4 lane de-

partures per 100 miles and reversal rates of 1 to 2 per minute - without any distracting tasks [27]. The measured number of steering wheel reversals falls within this range. However, measured number of lane departures in our study was approximately twice as high - regardless of condition and with a mean distance of only 11.81 km, meaning that participants drove serpentine-like but without abruptly moving the steering wheel. On the one hand, it is important to note that the experiment was performed in a medium fidelity simulator with naturally unrealistic vehicle physics. On the other hand, this suggests that hands-free multiparty phone conversations are very demanding and impair driving performance. Another point that could have potentially impaired driving performance is that all participants wore 3D glasses all the time. Overall, we did not uncover any main or interaction effects on driving performance suggesting that the support systems did not negatively affect driving performance.

However, the S3D visualization was perceived more *dependable* (UEQ, Figure 4) than the NoV-condition, regardless of the audio condition. Further, we did not measure any driving performance impairments of the visualization. That means that participants felt more in control and perceived the UI as more secure and predictable (c.f. Laugwitz et al. [17]) when the system presented visual indicators of who is speaking compared to when they had to rely on audio information only. This is supported by the mentioned effect on the *Situational Stress* scale of the DALI. A more dependable system might induce less stress. It is likely that the more obvious visual support system made people feel more certain in who is speaking and by that, made it easier to follow the conversation. Also, acting as an ambient lighting cue, participants can perceive the information in their peripheral vision without taking the eyes of the road. The presented names further enhance the experience by adding meta information. This is confirmed by responses which indicate the necessity of research on UIs containing additional information like the role of the speaker, faces or avatars, and locations during international phone conferences.

### 5.1. Limitations

Data suggests that the sample size was too small to generalize results. Follow-up experiments with more participants are necessary to confirm our conclusions. We missed to ask participants how often they engage in multiparty conversations. Considering that these support systems enhance conference calls while driving, another study with a sample that is very likely to often engage in such calls (unlike students and university staff in our study) is necessary to confirm results of our exploratory study. The experimental design did not investigate the full spectrum of visualizations like HUD or windshield displays as well as traditional 2D displays. While S3D is not crucial for this study and acts only as a design element, we chose S3D to explore the application domain of this visualization technique, in particular for structuring information. Further research is necessary to put the results in perspective to traditional 2D visualizations.

It is important to note that mental workload can vary substantially due to the type of conversation. Involvement and

engagement is another factor that can influence workload. In our study, participants might not have cared much about the conversation which influences final workload results.

Participants reported simulator sickness scores ranging from significant to problematic symptoms according to Kennedy et al. [11]. Taking a closer look at the scores reveals that the symptoms "difficulty concentrating" ( $\sum_{n=0}^N = 49$ ) and "sweating" ( $\sum_{n=0}^N = 29$ ) were reported especially high among participants compared to the average scores of the other symptoms ( $\sum_{n=0}^N = 8.9$ ). The first one is likely due to exhaustion. Participants needed a high effort of attention and reported high situational stress especially because of the simulated job interview. The latter is most likely due to the warm laboratory environment without appropriate air conditioning but many projectors, workstations and loudspeakers.

We need to mention that the used spatial sound setup is different from a typical setup encountered in a car. Especially the acoustics in the laboratory and a common car lead to differences in perceived proximity of audio sources and by that, the virtual speakers. Hence, replicating our experiment in a real car might lead to different results.

Also, being a simulator study, vehicle physics as well as overall replication of real-world conditions is limited by technology. Hence, results can only be interpreted within the context of our simulation. The absence of a baseline drive without any conversational task makes it hard to specify the reason for the impaired driving performance. However, comparing our results with previous work suggests that doing a multiparty call - e.g. a job interview - can lead to impaired driving performance. Hence, our data indicate that it is not recommended to engage in a multiparty conversations for manual driving with SAE Level 1 and 2 automation - even with support systems like ours. We suggest exploring the proposed support systems for higher levels of automation, e.g. Level 3 where participants have to monitor the environment but do not have to engage while automation is enabled.

## 6. Conclusion

In this work, we investigated two audiovisual support systems and their potential to make multiparty conferencing while driving more pleasant: a stereoscopic visualization and an object-based spatial auralization. We found no positive - but also no negative - impact of object-based audio in this context. This might be due to the spatialized noise reproduction. We propose to conduct further experiments of multiparty conferencing in realistic acoustic environments to verify if it generally reduces the benefits of using spatialized speech reproduction. However, results of our study indicate that a visualization presenting information about the callers potentially reduces perceived stress and is likely to increase attractiveness. Participants further perceived the conditions with a supporting stereoscopic 3D visualization as more dependable compared to user interfaces without a vision-based support system.

Considering the positive aspects of our prototype paired with the absence of any impairments, an intelligent visualization

has a lot of potential to support drivers during conference calls. Our results can be the basis for a user interface that offers more security, predictability, and makes users feel more in control during such phone calls. In our fast-paced society, where workplace availability and flexibility but also constantly keeping in touch with friends and family is of utmost importance, such a system could be highly beneficial for user experience and safety.

Since the impact on driving performance of such systems, the applied display, and the application in (semi-)autonomous vehicles are important aspects in this research domain, follow-up experiments with slightly modified designs that integrate these factors are planned to investigate this areas further. Especially a test with people who regularly participate in conference calls, the comparison with other types of displays (HUD, windshield, and perspective 3D), and the application of in vehicles with Level 3 automation seem promising.

## 7. Acknowledgements

This work has partially been funded by the Free State of Thuringia, Germany (FKZ: TUI-I-01-14 and 2015-FE-9109).

## 8. References

- [1] AHRENS, J., GEIER, M., RAAKE, A., AND SCHLEGEL, C. Listening and conversational quality of spatial audio conferencing. In *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space* (Oct 2010).
- [2] BERKHOUT, A. J. A holographic approach to acoustic control. *J. Audio Eng. Soc* 36, 12 (1988), 977–995.
- [3] BROY, N., ANDRÉ, E., AND SCHMIDT, A. Is stereoscopic 3D a better choice for information representation in the car? In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '12* (2012).
- [4] BROY, N., GUO, M., SCHNEEGASS, S., PFLEGING, B., AND ALT, F. Introducing novel technologies in the car - Conducting a Real-World Study to Test 3D Dashboards. In *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '15* (New York, New York, USA, 2015), ACM Press, pp. 179–186.
- [5] FIELD, A. P., MILES, J., AND FIELD, Z. *Discovering statistics using R*. Sage, London, 2014.
- [6] FRÖHLICH, P., SACKL, A., TRÖSTERER, S., MESCHTSCHERJAKOV, A., DIAMOND, L., AND TSCHELIGI, M. Acceptance Factors for Future Workplaces in Highly Automated Trucks. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '18* (New York, New York, USA, 2018), ACM Press, pp. 129–136.
- [7] GOLDING, J. F. Motion sickness susceptibility questionnaire revised and its relationship to other forms of sickness. *Brain research bulletin* 47, 5 (nov 1998), 507–16.
- [8] HOLMAN, T. *Surround Sound* (Second Edition). 2nd ed. Focal Press, Oxford, 2008.
- [9] HOWARD, I. P., AND ROGERS, B. J. *Binocular Vision and Stereopsis*. Oxford University Press, New York, New York, USA, 2008.
- [10] INKPEN, K., HEGDE, R., CZERWINSKI, M., AND ZHANG, Z. Exploring spatialized audio & video for distributed conversations. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work - CSCW '10* (New York, New York, USA, 2010), ACM Press, p. 95.
- [11] KENNEDY, R. S., DREXLER, J. M., COMPTON, D. E., STANNEY, K. M., LANHAM, S., AND HARM, D. L. Configural scoring of simulator sickness, cybersickness and space adaptation syndrome: Similarities and differences?
- [12] KENNEDY, R. S., LANE, N. E., BERBAUM, K. S., AND LILIENTHAL, M. G. Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness. *The International Journal of Aviation Psychology* 3, 3 (jul 1993), 203–220.
- [13] KILGORE, R., CHIGNELL, M., AND SMITH, P. Spatialized Audioconferencing: What Are the Benefits? In *Proceedings of the 2003 Conference of the Centre for Advanced Studies on Collaborative Research* (2003), CASCON '03, IBM Press, pp. 135–144.
- [14] KUN, A. L., AND MEDENICA, Z. Video Call, or Not, That is the Question. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2012), CHI EA '12, ACM, pp. 1631–1636.
- [15] KUN, A. L., VAN DER MEULEN, H., AND JANSSEN, C. P. Calling While Driving: an Initial Experiment With Hololens. *PROCEEDINGS of the Ninth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design* (2017), 200–206.
- [16] KUN, A. L., VAN DER MEULEN, H., AND JANSSEN, C. P. Calling while Driving using Augmented Reality: Blessing or Curse? (accepted for publication. *Presence: Teleoperators and Virtual Environments* (2018), 1–28.
- [17] LAUGWITZ, B., HELD, T., AND SCHREPP, M. Construction and Evaluation of a User Experience Questionnaire. 2008, pp. 63–76.
- [18] LAURIER, E. Doing Office Work on the Motorway. *Theory, Culture & Society* 21, 5 (2004), 261–277.
- [19] LEUNG, S., CROFT, R. J., JACKSON, M. L., HOWARD, M. E., AND MCKENZIE, R. J. A Comparison of the Effect of Mobile Phone Use and Alcohol Consumption on Driving Simulation



- Performance. *Traffic Injury Prevention* 13, 6 (2012), 566–574.
- [20] LÖCKEN, A., MÜLLER, H., HEUTEN, W., AND BOLL, S. AmbiCar: Towards an in-vehicle ambient light display. *AutomotiveUI 2013*, October (2013), 107–108.
- [21] MAURER, M., CHRISTIAN, G., LENZ, B., AND WINNER, H. *Autonomous Driving - Technical, Legal and Social Aspects*. 2016.
- [22] MELCHIOR, F. *Investigations on spatial sound design based on measured room impulse responses*. PhD thesis, Delft University of Technology, Delft, Niederlande, 24.06.2011.
- [23] PAL, M., ROY, R., BASU, J., AND BEPARI, M. S. Blind source separation: A review and analysis. In *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)* (nov 2013), pp. 1–5.
- [24] PAUZIE, A. A method to assess the driver mental workload: The driving activity load index (DALI). *IET Intelligent Transport Systems* 2, 4 (2008), 315.
- [25] PERTERER, N., MOSER, C., MESCHTSCHERJAKOV, A., KRISCHKOWSKY, A., AND TSCHELIGI, M. Activities and Technology Usage while Driving: A Field Study with Private Short-Distance Car Commuters. *Proceedings of the Nordichi '16: the 9Th Nordic Conference on Human-Computer Interaction - Game Changing Design* (2016).
- [26] RAJAN, R., CHEN, C., AND SELKER, T. Considerate Audio MEdiating Oracle (CAMEO). In *Proceedings of the Designing Interactive Systems Conference on - DIS '12* (New York, New York, USA, 2012), ACM Press, p. 86.
- [27] SAE INTERNATIONAL. Operational Definitions of Driving Performance Measures and Statistics J2944\_201506, jun 2015.
- [28] SCHWARTZ, B. L., AND KRANTZ, J. H. *Random Dot Stereograms*, 2018.
- [29] SKOWRONEK, J., AND RAAKE, A. Investigating the effect of number of interlocutors on the quality of experience for multi-party audio conferencing. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, August (2011), 829–832.
- [30] SZCZERBA, J., AND HERSBERGER, R. The Use of Stereoscopic Depth in an Automotive Instrument Display: Evaluating User-Performance in Visual Search and Change Detection. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 58, 1 (sep 2014), 1184–1188.
- [31] WEIDNER, F., AND BROLL, W. Exploring Large Stereoscopic 3D Dashboards for Future Automotive User Interfaces. In *Proceedings of the 9th International Conference on Applied Human Factors and Ergonomics AHFE*. Springer, Cham, Orlando, FL, 2019, pp. 502–513.
- [32] WICKENS, C. D. Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science* 3, 2 (jan 2002), 159–177.
- [33] YAP, B. W., AND SIM, C. H. Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation* 81, 12 (2011), 2141–2155.
- [34] YEO, I.-K., AND JOHNSON, R. A. A New Family of Power Transformations to Improve Normality or Symmetry. *biometrika Biometrika* 87, 4 (2000), 954–959.



# Abstract Reviewed Paper at ICSA 2019

Presented by VDT.

## New Potential for Portable Audio with MEMS based Speakers

D. Beer<sup>1</sup>, T. Brocks<sup>1</sup>, J. Küller<sup>1</sup>, S. Strehle<sup>2</sup>, T. Koch

<sup>1</sup>*Fraunhofer Institute for Digital Media Technology IDMT, 98693 Ilmenau, Germany*

<sup>2</sup>*Technische Universität Ilmenau, 98693 Ilmenau, Germany*

### Abstract

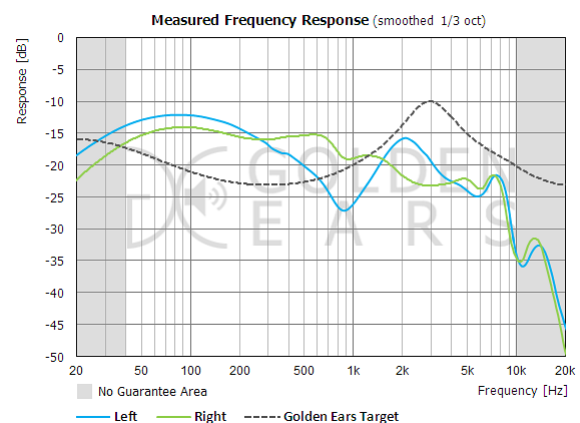
There is a high demand for portable audio on the market. Hence, manufactures of headphones and hearing aids have to deal with miniaturization of electroacoustic transducers (micro speakers) but maintain sound quality and energy efficiency (battery life time) at the same time. Different techniques are known for downsizing transducers. A very successful method is provided by the semiconductor industry. The use of the so-called MEMS technology (Micro-Electro-Mechanical-System) has already led to great success in microelectronics and MEMS applications such as microphones and accelerometers. This success has triggered a high interest in implementing MEMS technology also for speaker manufacturing. Based on patents, the initial approaches of MEMS loudspeakers will be presented. An outlook of the arising new potentials for portable spatial audio with MEMS based speakers will be given.

### 1. Motivation

3D audio is not only of high interest for loudspeaker reproduction, but also for headphones. Especially in the popular field of virtual reality (VR) and augmented reality (AR), headphones are the first choice. The connection to the ear is mostly fixed and every ear can directly be provided with audio content without any cross-talk. By the use of HRTFs (Head-Related Transfer Function) the virtual environment can be presented in a physical correct way [3, p. 9]. Knowing the exact sound pressure level of the headphones plays an important role in this application and could be much better achieved through the use of MEMS loudspeakers than classical approaches. In combination with a MEMS microphone this could be a perfect tool.

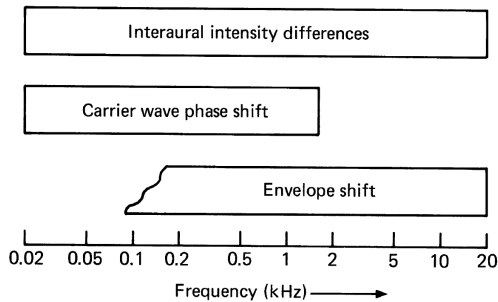
Other applications like active-noise-canceling (ANC) are also very popular. Nowadays hybrid feed-forward and feed-back filtering is used in mid to high prizing headphones. This means there is the need of at least two microphones per ear. When using in-ear headphones with small space requirements, the MEMS technology with all elements integrated in one chip can bring a big benefit.

In the manufacture of dynamic headphones derivations of several dB are common. Figure 1 shows the variation between left and right earphones in a dynamic in-ear headphone, which is not uncommon in the lower to mid price range.



**Fig. 1:** Frequency response of an earphone with deviations between left and right loudspeakers [8].

Big differences can cause a confusion because of inconsistent amplitude and phase of the audio signal [2, pp. 206–212]. In Fig. 2 the principle influence of phase and amplitude are shown in dependency of the frequency. Especially in the use for VR and AR, differences to the visual cues can be a problem. If the derivation is not too large, the brain can compensate this, but best way is to deliver an audio signal most accurate like a real sound event.



**Fig. 2:** Influence of various interaural advices on left and right hearing as a function of frequency [16, p. 639].

For higher pricing headphones the so called pairing (the two most similar drivers in one headphone) gives a better stereo representation. But the derivation from headphone to headphone is still there. Adaptive correction by microphone or signal processing can solve this problem. MEMS technology is ideal for this solution, as it allows a combination of high-precision MEMS loudspeakers with the smallest dimensions, MEMS microphones and digital signal processors (DSP) to be fabricated directly on a printed circuit boards (PCB).

## 2. MEMS Speakers

The following section explains MEMS technology in its basic principle and manufacturing process, including loudspeaker implementation. The historical and present approaches are presented.

### 2.1. MEMS Technology

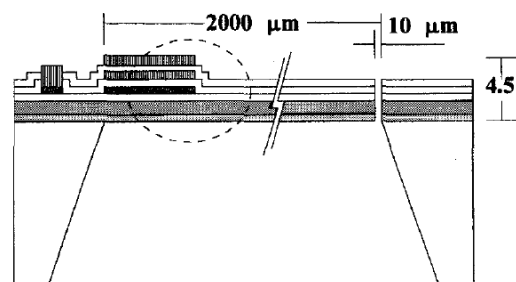
The abbreviation MEMS stands for **Micro-Electro-Mechanical System**. Hence, MEMS represent a rational combination of miniaturized components that operate typically in the mechanical and electrical domain. Based on the fact that semiconductor microelectronics and MEMS fabrication utilize in principle the same set of microtechnologies, enables (sub-)micrometer feature precision, high integration density, parallel high yield device manufacturing, high reproducibility and so-called monolithic device fabrication. Monolithic device fabrication means that all elements, comprising the micromechanical components and the required microelectronics are assembled in a miniaturized manner on the same substrate or at least within a single device package. This allows driving MEMS devices in principle with high energy efficiency and high performance signal processing. The packaged MEMS are relatively robust and furthermore allow a direct combination with conventional

PCB surface-mount technology. The monolithic integration combined with the aforementioned possibilities for high yield, high volume and parallel microdevice fabrication enables overall new design strategies for loudspeakers and potentially a device fabrication at relatively low costs. Nevertheless, the implementation of MEMS technologies also requires rethinking the overall speaker design. In contrast to a conventional speaker assembly, MEMS fabrication relies on the stacking of patterned functional layers (2.5D technology). These layers are created by additive and subtractive technologies comprising, for instance, vapor phase thin film deposition, optical lithography for micropattern transfer, and material dry as well as wet etching for area-selective material removal.

MEMS emerged from the developing microelectronics in the 60s. The so-called resonant gate transistor, demonstrated by Nathanson and Wickstrom [14], is typically considered as the first appearance of a MEMS device. In the 70s, pressure sensors and ink jet printer nozzles were developed based on MEMS principles [12, pp. 2] followed by the first MEMS based microphone in 1982 [17]. Since then, MEMS inertial sensors, microphones, magnetometers, micro-optical components and various other components were developed and represent today the foundation of modern sensor systems and communication devices.

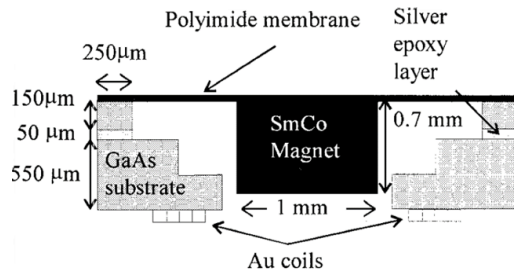
### 2.2. Early MEMS Speaker Approaches

The following overview is focused on MEMS based speakers for the reproduction of the audible frequency range. To the knowledge of the authors, the first approach of MEMS speakers was published in 1994 by Lee et al. [10] and was based on a piezoelectric cantilever. The concept itself was based on a sound generation by the stimulated bending of the microcantilever. Notably, the device was originally designed for microphone applications. A cross section of their speaker is shown in Fig. 3.



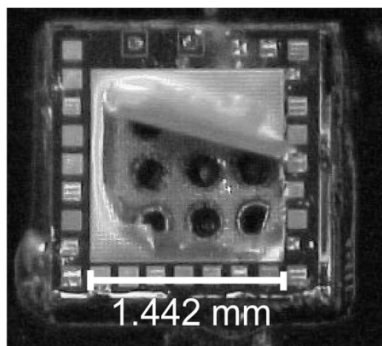
**Fig. 3:** Cross section of the piezoelectric MEMS-speaker from Lee et. al [10].

Only three years later Harradine et al. introduced the first electrodynamic MEMS speaker [6]. Attached to a membrane device, a permanent magnet interacts with a fixed voice coil, which caused generating of sound. A schematic illustration of this approach is shown in Fig. 4.



**Fig. 4:** Cross section showing the electrodynamic MEMS-speaker of Harradine et. al [6].

A MEMS speaker with electrostatic drive system was introduced by Neumann and Gabriel in 2002 [15]. Sound is generated by their system by the interaction of a movable membrane electrode and a fixed stator electrode. A photograph of their demonstrator is depicted in Fig. 5.



**Fig. 5:** Photograph of the electrostatic MEMS-speaker by Neumann et. al [15].

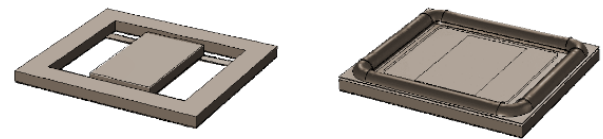
All recent MEMS speakers are more or less based on well known and well established principles from the field of common loudspeakers. Even exotic concepts like magnetostrictive drive systems are adapted to MEMS speaker approaches [1]. A more elaborate but still incomplete list of MEMS speaker related literature is given by Maennchen et al. [13]. In addition to these academic publications many patents and patent applications have been published by companies like Bosch [18], Infineon [4], Goertek [20] and AudioPixels [11]. In the meantime none of these companies have made any public demonstrations.

### 2.3. Today's MEMS Speakers

The following discourse is meant as short overview of the present MEMS speakers. A detailed view cannot be given in this frame. For further information on technical details or characterization results, please refer to the references provided. Due to the continuously growing demand for miniaturization of loudspeakers for headphone based applications, new approaches have been published in recent years:

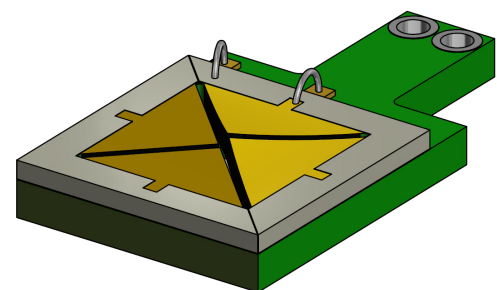
**USound** is an Austria-based startup and provides the only MEMS speaker available on the market at now. Multiple MEMS based cantilevers act as piezoelectric drive system. Attached to this is a plate that acts as piston-like membrane. Because the membrane is attached in a non-semiconductor-

technology processing step, it is a MEMS hybrid [19]. The MEMS based drive system and the entire MEMS speaker is depicted in Fig. 6.



**Fig. 6:** Schematic illustration of the USound loudspeaker [19].

**Fraunhofer ISIT** introduced a piezoelectric MEMS speaker consisting of four triangular cantilevers which are separated by narrow gaps. This speaker is hence called piezoelectric narrow gap MEMS (PNG-MEMS) [13]. Related to the drive signal, the cantilevers act as bending actuators for the reproduction of sound. Because of the narrow gaps, the acoustic short-circuit is avoided and large excursions are enabled. Functionality was proven by public demonstrations of an In-Ear Headphone demonstrator several times, at first at the Conference of the German Acoustic Society DAGA in 2018. In order to optimize its performance, the In-Ear Headphone was designed and equipped with dedicated signal processing by Fraunhofer IDMT. The basic principle of the PNG-MEMS is shown in Fig. 7.



**Fig. 7:** Working principle of the ISIT loudspeaker [13].

**Fraunhofer IPMS** developed an electrostatic MEMS speaker with a drive system called Nano-Electrostatic-Drive (NED). The working principle differs from that of conventional electrostatic speakers and is based on electrostatic actuated bimorph bending actuators moving in-plane [5, 9]. Within the chip, several actuators move in pairs towards and away from each other. In that way air is pushed out of one speaker side and sucked in on the other. Some experts may recall the Air Motion Transformer from Oskar Heil [7]. Functionality was proven by demonstrations of an In-Ear Headphone in the project team of Fraunhofer IPMS and Fraunhofer IDMT. The basic principle of the NED is shown in Fig. 8.

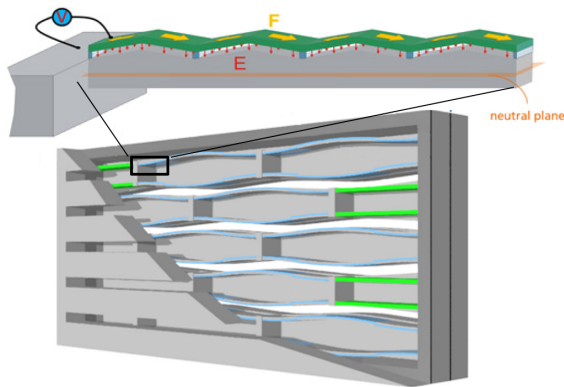


Fig. 8: Working principle of the IPMS loudspeaker [5].

### 3. Outlook

The first developments of MEMS loudspeakers presented in this paper provide high potential for headphone based applications. Further considerations regarding performance and new actuator designs are required.

#### 3.1. Conclusion

More than any other technology, MEMS technologies provide a high potential to fulfill the current need and to effectively address the current challenge of miniaturized speakers considering high acoustical performance and low prices for portable audio applications like headphones, hearables, hearing aids, and AR respectively VR glasses. The technology ensures simple, high volume production of speakers, which are more light weighted, less divergent and easier to assemble than conventional technologies. Especially the high integration density of actuators, sensors and signal processing and the high-precision manufacturing make MEMS loudspeakers highly attractive. All components can be easily interconnected and adjusted to each other. This enables sufficient versatility to cope with various application scenarios. Regarding headphone system functions like adaptive control to compensate deviations between left and right channels, automated installation of user-oriented listening experiences and similar are possible.

#### 3.2. Future Work

MEMS speaker technology follows a paradigm change in the value chain of headphone based solutions. As a result, headphones are not manufactured anymore by a single manufacturer, but represent a combination of headphone manufacturers, chip manufacturers, and design and software developers. The increase in production can also reduce manufacturing costs. As more and more systems are being developed by wireless means, the problem with impedance loads is becoming irrelevant, which means that the power consumption of mobile devices is reduced. Further development will prove to what extent MEMS loudspeakers can also be used as mobile loudspeakers in mobile devices, as headphones with over-ear or even as free-field loudspeakers.

## 4. References

- [1] Thorsten S. Albach and Reinhard Lerch. 2013. "Magnetostrictive microelectromechanical loudspeaker". *The Journal of the Acoustical Society of America* 134, 6 (2013), pp. 4372–4380.
- [2] J. Blauert. 1974. *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press. ISBN:0262024136.
- [3] J. Breebaart and C. Faller. 2008. *Spatial Audio Processing: MPEG Surround and Other Applications*. Wiley. ISBN:9780470723487.
- [4] Alfons Dehe and Shu-Ting Hsu. 2013. "Schallwandler mit einer ersten und einer zweiten Menge von ineinandergreifenden Kammfingern". Patent DE 10 2012 220 819 A1.
- [5] Lutz Ehrig, Bert Kaiser, Hermann A. G. Schenk, Michael Stolz, Sergiu Langa, Holger Conrad, Harald Schenk, Andreas Maennchen, and Tobias Brocks. 2019. "Acoustic Validation of Electrostatic All-Silicon MEMS-Speakers". In *Audio Engineering Society Conference: 2019 AES International Conference on Headphone Technology, San Francisco*.
- [6] MA Harradine, TS Birch, JC Stevens, and C Shearwood. 1997. "A micro-machined loudspeaker for the hearing impaired". In *Proceedings of International Solid State Sensors and Actuators Conference (Transducers '97), Chicago*. IEEE, pp. 429–432.
- [7] Oskar Heil. 1972. "Acoustic transducer with a diaphragm forming a plurality of adjacent narrow air spaces open only at one side with the open sides of adjacent air spaces alternatingly facing in opposite directions". Patent US 3636278 A.
- [8] Golden Ears. "Measurements BOSE IE2". (accessed 26.09.2019). [http://en.goldenears.net/index.php?mid=GR\\_Earphones&page=11&document.srl=13633](http://en.goldenears.net/index.php?mid=GR_Earphones&page=11&document.srl=13633)
- [9] B. Kaiser, L. Langa, S. and Ehrig, M. Stolz, H. Schenk, H. Conrad, H. Schenk, K. Schimmanz, and D Schuffenhauer. 2019. "Concept and proof for an all-silicon MEMS micro speaker utilizing air chambers". *Nature Microsystems and Nanoengineering* 5, 43 (2019), pp. 1–11.
- [10] Seung S. Lee, R. P. Ried, and R. M. White. 1994. "Piezoelectric Cantilever Microphone and Microspeaker". *Solid-State Sensor and Actuator Workshop Hilton Head, South Carolina* 5, 4 (1994), pp. 238–242.
- [11] Audio Pixels Limited. (accessed 26.09.2019). <http://www.audiopixels.com.au/index.cfm/audio-pixels/>
- [12] C. Liu. 2012. *Foundations of MEMS*. Prentice Hall, Chapter 1. ISBN:9780132497367.

- [13] Andreas Maennchen, Fabian Stoppel, Tobias Brocks, Florian Niekiehl, Daniel Beer, and Bernhard Wagner. 2019. "Design and Electroacoustic Analysis of a Piezoelectric MEMS In-Ear Headphone". In *Audio Engineering Society Conference: 2019 AES International Conference on Headphone Technology, San Francisco*.
- [14] H.C. Nathanson. 1965. "Resonant gate transistor". Patent US 3413573.
- [15] J. J. Neumann and K. J. Gabriel. 2002. "CMOS-MEMS membrane for audio-frequency acoustic actuation". In *Technical Digest. MEMS 2001. 14th IEEE International Conference on Micro Electro Mechanical Systems, Pittsburgh*. pp. 236–239.
- [16] C. Poldy. 2001. *Loudspeaker and Headphone Handbook*. Focal Press, Oxford, Chapter 14 - Headphones, pp. 585–692. ISBN:9780240522760.
- [17] M. Royer. 1982. "Piezoelectric pressure sensor". Patent US 4445384 A.
- [18] Christoph Schelling and Thomas Northemann. 2017. "MEMS-Lautsprechervorrichtung sowie entsprechendes Herstellungsverfahren". Patent DE 10 2016 201 872 A1.
- [19] USound. "Technology". (accessed 26.09.2019). <https://www.usound.com/technology/>
- [20] Qunbo Zou and Zhe Wang. 2018. "Mems device and electronics apparatus". Patent WO 2018/064S04 A1.







## Full Reviewed Paper at ICSA 2019

Presented\* by VDT.

### Switched Spatial Impulse Response Convolution as an Ambisonic Distance-Panning Function

Patrick Cairns<sup>1</sup>, Dr David Moore<sup>2</sup>

<sup>1</sup> Glasgow Caledonian University, Glasgow, Scotland, Email: patrickcairns1991@gmail.com

<sup>2</sup> Glasgow Caledonian University, Glasgow, Scotland, Email: j.d.moore@gcu.ac.uk

#### Abstract

*Ambisonics offers a robust and effective approach to the recording, processing and delivery of Spatial Audio. The Ambisonic system is often considered to provide a perceptually and computationally advantageous Spatial Audio experience in comparison to typical Binaural systems. This is true even when an end-step Binaural render is required, as is typical in Virtual or Augmented Reality systems which naturally imply audio delivery via headphones.*

*Standard Ambisonic processing allows for the rotation of a sound field around an origin position. There is not, however, a strongly established means of modulating the radial distance of a virtual sound source from the origin.*

*This paper presents a potential solution to an Ambisonic distance-panning function for both static and dynamic virtual sources in the form of a FOA (First Order Ambisonics) Switched-SIR (Spatial Impulse Response) Convolution Reverberator. This includes a presentation of the framework for such a function, and an analysis of audio rendered using prototype scripts.*

#### 1. Introduction

Two prominent approaches to Spatial Audio have become popular in applications: Binaural Audio, and Ambisonics (though less used approaches such as Wave Field Synthesis are also available) [1], [2].

Ambisonics is generally considered to be an accurate and perceptually satisfying depiction of Spatial Sound, even when considered in systems where Binaural Audio rendering is required at the end step for headphone delivery [2], [3], [4], [5], [6], [7]. The principles of Ambisonics allow for an optimised Spatial Audio processing medium which is being enthusiastically adopted by the cutting edge of the audio industry.

Typical Ambisonic processing only allows for the rotation of a 3-dimensional sound field around a centre point, providing no control over sound source distance from this origin [8].

Though a degree of research has been undertaken to provide control over the distance parameter in Ambisonic processing there remains no uniform method of accomplishing this

function. Current designs largely fall into one of three categories:

1. Systems which that are only typically appropriate for modelling the free field (Wave Field Synthesis, amplitude attenuation) [9], [10], [11].
2. Systems which allow modulation of listener position within a sound-field but not the modulation of the radial distance between a virtual source and sound field centre (Virtual Loudspeaker approach) [5], [6].
3. Reverberators which are only appropriate for accurately modelling regular rectangular rooms or diffuse fields [12], only model static sound sources [13], or are only viable up to a small order of reflections in real-time application (Geometric Simulation) [14], [15], [16].

It can be clearly seen that such designs, though innovative and useful, do not meet the criteria required of an Ambisonic distance panning function for modern Spatial Audio applications: The ability to accurately place or emulate the

placement of sound sources in a sound field within a complex acoustic environment, the ability to render dynamic audio over varying distance, and viability in real-time application for dynamic systems.

This paper offers a solution to an Ambisonic distance-panning function that meets these criteria in the form of Switched SIR (Spatial Impulse Response) Convolution. In this system SIR sets describing a range of discrete distances for a specified acoustic environment may be convolved with a mono input signal to provide an Ambisonic Auralisation at specified distances. By ‘switching’ the SIR set being convolved it is possible to create Ambisonic Auralisations of dynamic sound sources moving across distance. An architecture for this solution is presented here, including an overview of the system development and design, and an assessment of audio rendered using prototype functions with relevance to the success and viability of the design (namely time-frequency analysis and listening tests).

## 2. Background

### 2.1. Ambisonics

Ambisonics is a system of full periphonic directional sound pickup, storage, processing and reproduction developed through the 1970’s by Gerzon, Fellgett and Barton among others, taking influence from the earlier work of Cooper and Shiga [1], [17].

Ambisonics describes a sound field around an origin position, and with radius equal to the radial distance of a sound source from this origin position, using the spherical expansion of the wave equation in the form [18], [19], [20]:

$$p(\mathbf{r}, \theta, \phi, k) = \sum_{n=0}^{\infty} \sum_{m=-n}^n A_n^m(k) j_n(kr) Y_n^m(\theta, \phi) \quad (2.1.1)$$

Where  $p(\mathbf{r}, \theta, \phi, k)$  is the pressure at a point in space,  $k$  is the wave number,  $r$  is the radial distance,  $\theta$  is the elevation, and  $\phi$  is the azimuth.

The Spherical Harmonics,  $Y_n^m(\theta, \phi)$ , and Spherical Bessel Function of the first kind,  $j_n(kr)$ , describe a unit sphere in terms of functions on the surface of the sphere and radial functions respectively where  $n$  is the order and  $m$  is the degree of the Spherical Expansion.

The Spherical Harmonics are given as [18], [21]:

$$Y_n^m(\theta, \phi) = N_n^m P_n^m(\cos\theta) e^{jm\phi} \quad (2.1.2)$$

Where  $P_n^m(\cos\theta)$  is the Legendre function, describing angle of elevation, and  $e^{jm\phi}$  describes the azimuth.  $N_n^m$  is a normalisation factor, typically given as the SN3D normalisation scheme [18], [21], [22]:

$$N_n^m = \sqrt{\frac{(2n+1)(n-m)!}{4\pi(n+m)!}} \quad (2.1.3)$$

The Spherical Bessel Function of the first kind is given as:

$$j_n(x) = (-1)^n x^n \left(\frac{1}{x} \frac{\partial}{\partial x}\right)^n \frac{\sin(x)}{x} \quad (2.1.4)$$

Where in the system presented  $x=kr$  thus describing the radial functions [19], [20].

The Ambisonic Coefficients,  $A_n^m(\mathbf{k})$ , describe the content of the sound field, and, using the free-field spherical decomposition, may be given as [18], [21]:

$$A_n^m(k) = \frac{1}{j_n(kr)} \int_0^{2\pi} \int_0^\pi p(r, \theta, \phi, k) Y_n^m(\theta, \phi)^* \sin(\theta) d\theta d\phi \quad (2.1.4)$$

Each discrete order and degree of Ambisonic Coefficient corresponds to an audio channel in the Ambisonic B-Format. In First Order Ambisonics (FOA) the four B-Format audio channels describe a sound pressure in terms of an omnidirectional pressure component and plane waves along each of the three orthogonal spatial dimensions, with each component designated a discrete B-Format channel. The nomenclature for these B-Format channels has classically followed the Furse-Malham (FuMa) scheme, though more recent systems show a rising popularity in the use of the Ambisonics Exchangeable (AmbiX) scheme [23].

n	m	AmbiX ACN	FuMa Channel	$A_n^m$
0	0	0	W	1
1	-1	1	Y	$\sin\theta \cos\phi$
1	0	2	Z	$\sin\theta$
1	1	3	X	$\cos\theta \cos\phi$

**Tab. 2.1.1:** First Order Ambisonic Components.

### 2.2. Spatial Impulse Response

Impulse Response (IR) is a function describing the filtering effect of any system considering the output with respect to the input, and may be described mathematically in the convolution equation [24]:

$$y(n) = x(n) \otimes h(n) = \sum_{k=-\infty}^{\infty} x(n-k)h(k) \quad (2.2.1)$$

Here  $x(n)$  is the filter input,  $y(n)$  is the filter output and  $h(n)$  denotes the impulse response.

In audio applications IR is widely used to describe the acoustic influence of a space on any sound actualised in the space. IR may be measured in acoustic spaces [25], [26], estimated using statistical approaches [27], or rendered through geometric simulation [16], [28]. IR measurement procedure considers an excitation signal as the system input and the recorded or simulated output (typically from an omnidirectional microphone) as the output. The excitation signal used is required to provide an even signal across the time and frequency domain such that the acoustic properties of the room may be measured evenly. Several approaches are available for providing this excitation signal such as MLS or Sine Sweep method [25].

The Logarithmic Sine Sweep signal, as used in this paper, takes the form [25], [29]:

$$s(t) = \sin\left[Ke^{-\frac{t}{L}} - 1\right] \quad (2.2.2)$$

Where  $s(t)$  is the excitation signal,  $t$  is time and  $K$  and  $L$  are given as:

$$K = \frac{\omega_1 T}{\ln\left(\frac{\omega_1}{\omega_2}\right)} \quad (2.2.3)$$

$$L = \frac{T}{\ln\left(\frac{\omega_1}{\omega_2}\right)} \quad (2.2.4)$$

Where  $T$  is the duration of the sweep, and  $\omega_1$  and  $\omega_2$  are the start and end frequencies for the sweep respectively.

The impulse response is obtained for the sine sweep method by creating an inverse filter,  $f(t)$ , from the excitation signal,  $s(t)$ , and convolving with the room response to the excitation signal,  $r(t)$ , in order to obtain the linear impulse response,  $h(t)$ :

$$h(t) = r(t) \otimes f(t) \quad (2.2.5)$$

The inverse filter,  $f(t)$ , is calculated by reversing the excitation signal and then applying an amplitude modulation filter of +6dB per octave. This modulation filter may be given as:

$$m(t) = \frac{\omega_1}{\omega(t)} \quad (2.2.6)$$

Where the introduced term,  $\omega(t)$ , is the instantaneous frequency for each sample of  $t$ .

Spatial Impulse Response (SIR), also sometimes referred to as Directional Room Impulse Response (DRIR), differs in that the output of the system is the B-Format audio channels rather than the mono audio channel delivered by standard IR measurement [30], [31]. As such SIR is given in IR sets describing the difference between a mono excitation signal and each of the B-Format channels.

It should be noted that with mono-Ambisonic upmixing widely available it is a simple task to render an excitation signal as B-Format channels and obtain SIR sets describing the input-output difference for discrete Ambisonic components (B-format to B-Format rather than Mono to B-Format).

SIR may be measured using B-Format microphones, or rendered through the placement of virtual B-Format transducer arrays in geometric simulation. The obtained SIR may be applied to audio in order to render that audio in the measured acoustic environment as an Ambisonic sound field with great accuracy by convolving the SIR set with audio signal input. This process is described as Ambisonic Auralisation [30], [32].

### 2.3. Real-Time Convolution

Direct convolution of an impulse response and audio file is a computationally expensive method of rendering audio. The amount of data that is required to be processed in a short space of time in audio applications requires a further set of fast-convolution techniques [33].

Computational cost may be saved using Fast Fourier Transform (FFT) methods to compute the Transforms, achieve convolution through multiplication in the frequency domain, and compute the Inverse Fast Fourier Transform (IFFT) to give the convolved output [24]. This FFT method is generally based on the ‘Cooley-Tukey algorithm’, or ‘Radix-2 decimation’ [34], [35].

If the Fourier transform is given for an N-term sequence as:

$$F[k] = \sum_{n=0}^{N-1} f[n]e^{-2jnk\pi/N} \quad (2.3.1)$$

Then by declaring a variable called the ‘twiddle factor’, defined as:

$$W = e^{-2j\pi/N} \quad (2.3.2)$$

Substituting in the ‘twiddle factor’:

$$F[k] = \sum_{n=0}^{N-1} f[n]W^{nk} \quad (2.3.3)$$

Where  $W$  is constant for a fixed value of  $N$ . It is then possible to take advantage of the properties of symmetry and periodicity to drastically decrease the number of computations required.

$$\text{Symmetry: } W_N^{k[N-n]} = W_N^{-kn} = (W_N^{kn})^* \quad (2.3.4)$$

$$\text{Periodicity: } W_N^{kn} = W_N^{k[n+1]} = W_N^{[k+N]n} \quad (2.3.5)$$

Breaking down the transform this way decreases the amount of computations required from a factor of  $N^2$  for the Discrete Fourier Transform to a factor of  $N \log N$  [34].

Partitioned convolution algorithms break down the input signal and impulse response into blocks of samples which may be convolved in real-time using FFT algorithms [35]. The most common partitioning scheme is the Overlap-Add method, where the output of each block of convolution is summed into the system output at the relevant sample index as defined by the partitioning algorithm [33].

The overlap-add process can be outlined as [33]:

1. Partition the input signal into segments
2. Zero Pad the input blocks and impulse response to an equal and even length of FFT.
3. For each zero-padded input segment perform the FFT, Frequency Domain multiplication and IFFT.
4. For each resultant block sum into the output from the relevant sample.

## 3. System Overview

The Ambisonic distance-panning function offered in this paper uses SIR convolution to create Ambisonic Auralisations of a mono input signal. SIR sets, where each SIR describes a discrete distance, may be obtained through acoustic measurement or simulation. Each discrete SIR in such an obtained set describes an impulse function in terms of a spherical sound field of specified radial distance. Convolution of a mono sound source with any discrete SIR therefore renders the input as a virtual source at the relevant radial distance.

In typical overlap-add partitioned convolution algorithms a single IR is called for convolution each time a new block of the input signal is partitioned [33]. In the switched-SIR system presented here the SIR called for convolution can be varied at each partition. If we consider the instance where a sequence of SIRs are called which describe a location A and progress sequentially towards location B then the resultant Ambisonic Auralisation can be considered an emulation of a dynamic sound source.

### 3.1. Practical Considerations

The system developed only extends to First Order Ambisonics (FOA), largely due to the availability of equipment. It can be

recognised that the design presented may indeed be extended to include higher order Ambisonic systems [13].

The system receives mono input, as this is the most common input form for Auralisation purposes [8]. It can, however, be easily seen that through various upmixing capabilities SIR sets for other input formats may be easily computed.

### 3.2. Previous Work

The Switched-SIR convolution design draws largely from two key systems:

1. The convolution panning system presented by Stewart and Sandler [36]. This system presents the key framework for switched-IR convolution as a panning function, and in turn takes roots in the head-tracking system developed by Reilly and McGrath which uses the same functionality with HRTFs [37].
2. The Ambisonic convolution architecture presented by Lopez-Lezanco [13] which provides the framework for the application of SIR convolution.

## 4. System Design

### 4.1. SIR Measurement

In order to obtain an SIR set which describes the distance dimension a set of real SIRs were measured in Lecture Theatre W011 at Glasgow Caledonian University.

The room was in an unoccupied state, and the noise floor was recorded at 40dB. A Dodecahedral Loudspeaker calibrated at 80dB (not ideal but more importantly within safe listening levels) was set up and used to output the excitation signal. The excitation signal used was generated in Reaper using the ReaVerb plugin to provide a 1.5s Logarithmic Sine Sweep.

A SoundField ST250 microphone was used to record the B-Format room response to the excitation signal at discrete source-receiver distances, varying from 1m to 12m at 0.5m intervals. Both source and receiver were set at 1.5m height, taking care not to vary azimuth or elevation as distance was varied. The ReaVerb plugin was then used to deconvolve the SIR on a channel-by-channel basis and SIRs were stored as .wav files.

At each discrete SIR measurement distance a B-Format recording is made of a semi-anechoic speech sample in order to provide reference material for system analysis.

### 4.2. Rendering Audio

A set of prototype functions were developed using MATLAB to render audio to assess the Switched-SIR distance panning design using the recorded SIR set describing discrete points along the distance dimension. A mono semi-anechoic speech sample was selected as the input signal for providing a range of static and dynamic Ambisonic Auralisations. The computer used to run these scripts features a 4GHz quad-core processor and 16GB of DDR4 RAM.

#### 4.2.1. Static Render

The MATLAB script used to provide the static Ambisonic Auralisations consisted of a simple convolution using the Radix-2 Decimation FFT algorithm.

The mono speech sample and SIR at a manually specified distance are loaded by the script. The SIR being convolved is zero-padded to the length of the speech sample. FFTs are performed on each channel of the SIR and on the mono input signal. The mono input signal is then convolved with each channel of the SIR through multiplication in the Fourier (frequency) domain. IFFTs are then taken for each of the resultant outputs, providing the FOA B-Format output channels containing the static Ambisonic Auralisation.

This script was used to render B-Format audio at distances of 1m, 2m, 4m, 8.5m and 12m.

This process was measured as taking 0.196037s to process an excess of 2 million samples, just under 0.1 microseconds per sample, and again clearly showing that real-time application is viable with the implementation of partitioned convolution algorithms, given that at 44.1KHz we are passing 1 sample every 22.7 microseconds, this shows that block sizes of around 200 samples are a viable option for partitioning of four channels.

#### 4.2.2. Dynamic Render

The MATLAB script used to provide the dynamic renders introduces the switched-SIR partitioning scheme in timed offline renders.

Firstly the complete SIR set describing the distance dimension and the mono speech sample are loaded by the script. The SIR set is then zero padded to equal length and FFTs are performed on each SIR. In application the pre-computing of SIR FFTs is sensible to save computing cost during online application.

The dynamic renders are specified for a dynamic sound source moving at a speed of 1m/s over 4.5m to maintain consistency for analysis. As such the input signal is partitioned at 0.5s intervals. Each partition is convolved with a discrete SIR, switching the SIR at each partition to move sequentially along the distance dimension. These convolutions are performed using the same FFT algorithm as with the static renders, and implement the overlap-add method to sum each partitioned block into the output B-Format channels.

This script was used to render dynamic sound sources travelling across distance from 1m to 5.5m, 4m to 8.5m, 7m to 11.5m, 6m to 1.5m, 9m to 4.5m, and 12m to 7.5m.

This section can be measured as taking 0.880042 seconds to complete the convolution and Inverse Fourier Transform for over 40 million samples, indicating once again that real-time application of such a system is easily viable with real-time partitioned convolution techniques given that this time can be reduced using more elegant algorithms.

#### 4.2.3. Binaural Render

In order to provide a point of comparison with typical Spatial Audio distance-panning technology equivalent-distance static and dynamic Binaural Renders were created using the mono speech sample and a typical Binaural processing tool, which allows distance modulation through simple amplitude attenuation and the computation of reflections up to the 3<sup>rd</sup> order.

## 5. System Analysis

The success and viability of the developed Switched-SIR Ambisonic distance panning function for rendering static and dynamic B-Format audio was assessed through Time-Frequency Analysis and Listening Tests.

### 5.1. Time-Frequency Analysis

#### 5.1.1. Static Ambisonic Renders

Analysis of static Ambisonic renders was conducted in order to examine similarity to the reference recordings made during the SIR measurement process.

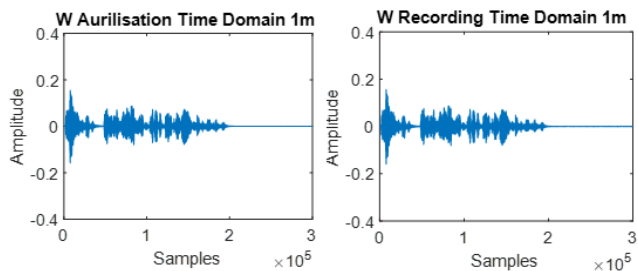


Fig. 5.1.1.1: Aurilised Vs Recorded B-Format W channel at 1m.

Observing in the time domain it is evident how accurate an Aurilisation SIR Convolution provides even when using FFT algorithms.

Peak-matching between the Aurilisation and recordings consistently showed only miniscule differences between iterations of each audio channel.

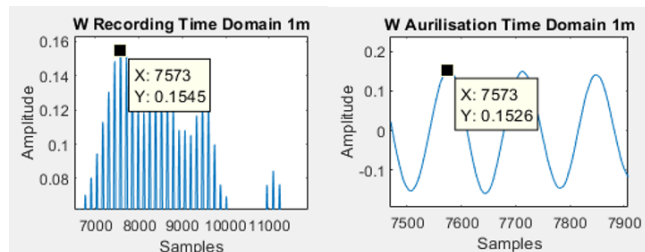


Fig. 5.1.1.2: Peak-matching between recordings and static Aurilisation at 1m.

Spectral comparison between the static Aurilisations and recording did however show certain inconsistencies between the two.

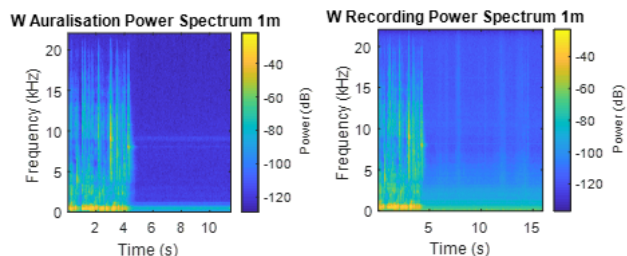


Fig. 5.1.1.3: Power Spectrum of recordings and static Aurilisation at 1m.

Though extremely accurate the Aurilisations notably contain more ‘defined’ spectral qualities (showing a lessening of the spectral smearing associated with reverberation) and

contained less late high-frequency energy, presumably due to the length of the SIRs.

The smearing at around 8-9kHz and mains hum apparent in the Aurilisations are simply the effects of non-ideal SIR measurement in a busy city centre campus.

#### 5.1.2. Dynamic Ambisonic Renders

As no options were available for providing dynamic Ambisonic recordings of the mono speech sample the dynamic Ambisonic Aurilisations are rendered as Binaural Stereo and evaluated with comparison to the dynamic Binaural renders created using the typical Binaural processing tool.

In the time domain the Ambisonic Aurilisation can be seen as an accurate render when viewed side-by-side to the Binaural-only processing equivalent, indicating that the Switched-SIR method does indeed provide a valid means of rendering dynamic sound sources. The Ambisonic Aurilisation can also be seen as providing more natural tails than the Binaural processing where the reverberant levels quickly fall due to the low order of computed reflections.

This is also visible in the waterfalls plots.

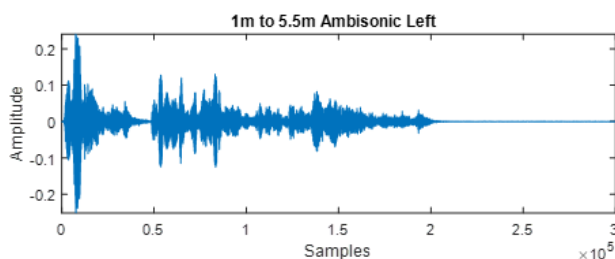


Fig. 5.1.2.1: Dynamic Ambisonic Aurilisation from 1m to 5.5m in time domain, left Binaural channel

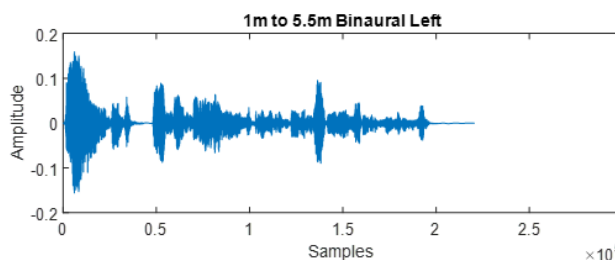


Fig. 5.1.2.2: Dynamic Binaural Render from 1m to 5.5m in time domain, left Binaural channel.

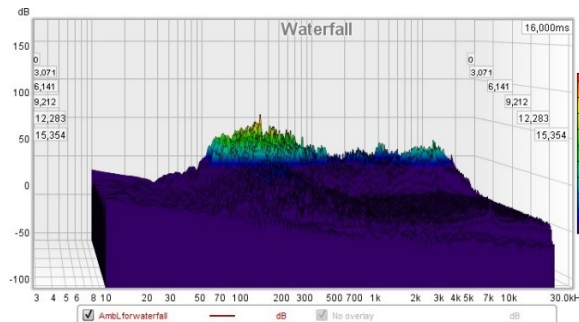


Fig. 5.1.2.3: Left Channel Waterfall of Binaural End-Step Render of Dynamic Ambisonic Aurilisation moving from 7m to 11.5m.



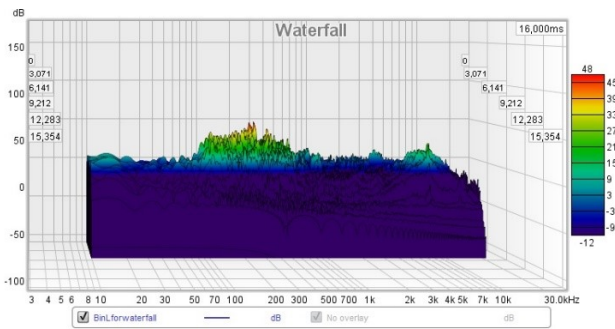


Fig. 5.1.2.4: Left Channel Waterfall of Binaural process Render of Dynamic Sound Source moving from 7m to 11.5m.

## 5.2. Listening Tests

Listening tests were conducted to assess the ability of the developed system to accurately deliver auditory distance cues for both static and dynamic renders.

As a full Ambisonic playback system was impractical, and again end-step Binaural renders are a typical component in Spatial Audio systems, audio for these listening tests were rendered as Binaural Stereo. This allows an assessment of the switched SIR Ambisonic distance panning function as a processing option for Binaural delivery in contrast to typical Binaural distance processing.

As absolute source-receiver distance is notoriously inaccurately perceived using only auditory cues where a user has no other indication to their environment a more meaningful measure of distance perception was utilised: relative distance between two sound sources.

### 5.2.1. Static Renders

13 experienced listeners were asked to indicate the absolute perceived distance of a static sound source on a continuous numerical scale. Static renders at 1m, 2m, 4m, 8.5m and 12m were played via headphones in random order. Using the first indicated perceived distance as a ‘reference’ measurement it was then possible to extract information on the percentage error of accuracy in relative distances perceived between sound sources rendered at varying distance.

This process was conducted for static Ambisonic Auralisation, static Ambisonic Recordings and static Binaural-only processing, and results were subjected to a one-way Analysis of Variance (ANOVA).

Static Ambisonic Recordings indicated that the accuracy of perceived relative distance does indeed vary significantly over a range of distances. The significant outlier can be clearly identified as the 3m discrepancy, this is likely due to the nature of percentage errors, and indicates some testing redesign may be in order.

ANOVA of results from the Static Ambisonic Auralisations indicated that relative distance can indeed be perceived with consistent accuracy over varying distances, though again the largest percentage error can be seen in smaller discrepancies. Otherwise results exhibited a similarity to the results from the Ambisonic recordings. The increased accuracy of the Ambisonic Auralisation compared to Ambisonic recordings can presumably be attributed to the reduced late reverberant

energy, thus reducing the natural smearing effect of reverberation on sound source localisation [38], [39].

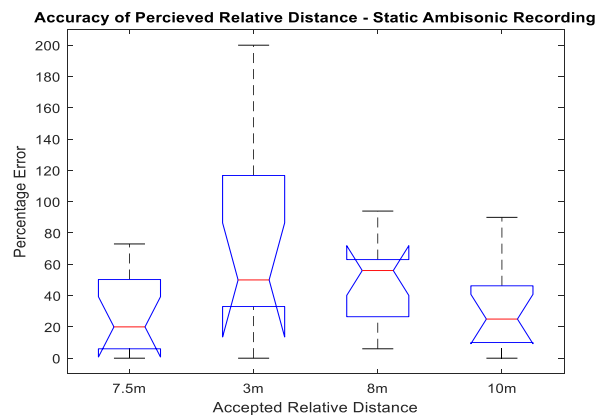


Fig. 5.2.1.1: Notch Graph of results from Static Ambisonic Recording listening tests.

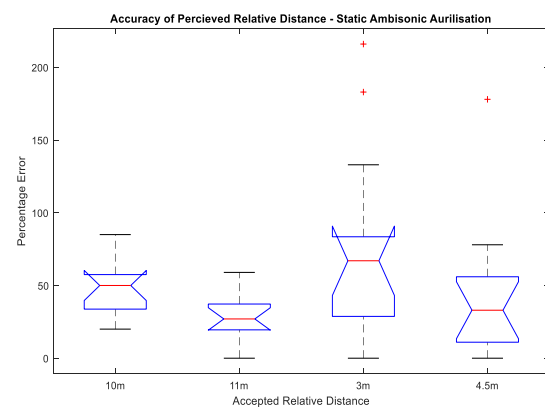


Fig. 5.2.1.2: Notch Graph of results from Static Ambisonic Auralisation listening tests.

The Binaural-processing render listening test results showed another significant outlier at 3m, indicating relative distance is not perceived consistently as distance varies, but did however otherwise show greater accuracy than expected from natural listening conditions. This could again suggest that the lack of natural reverberation in the rendering provides unnaturally accurate localisation [38], [39].

### 5.2.2. Dynamic Renders

13 experienced listeners were asked to indicate the start and stop distance of a dynamic sound source played over headphones on a continuous numerical scale. Sound sources moved over distances of 1m to 5.5m, 4m to 8.5m, 7m to 11.5m, 6m to 1.5m, 9m to 4.5m, and 12m to 7.5m. This test was undertaken under the same form for the audio rendered from Binaural-only processing and from the Switched SIR Ambisonic Distance panning function presented in this paper.

One-way ANOVA of results showed that neither approach yielded consistent perception of relative distance over varying distances. The results were, however, extremely similar, indicating the Switched SIR system performs comparably to the Binaural system in this regard. The inconsistency is considered as possibly due to the overall poor performance of

the human auditory system without accompanying perceptual cues, and that these results mirror these expectations from natural listening conditions [38] [39].

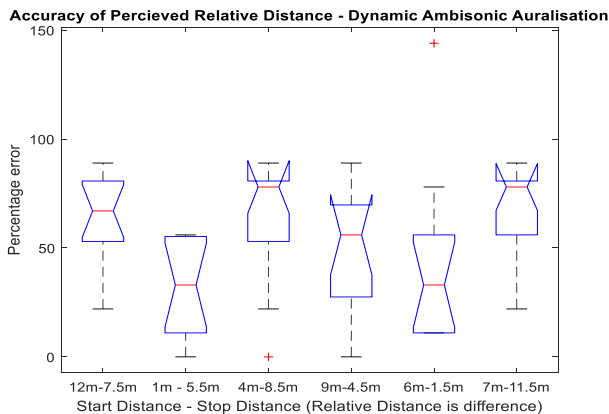


Fig. 5.2.2.1: Notch Graph of results from Dynamic Ambisonic Auralisation listening tests.

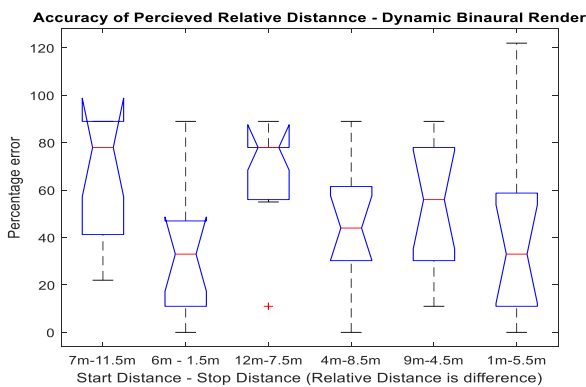


Fig. 5.2.2.2: Notch Graph of results from Dynamic Binaural render listening tests.

## 6. Conclusion

Time-frequency analysis of static renders confirms that Ambisonic Auralisation is an effective method of rendering a mono input as an Ambisonic sound source across distance. Results would indicate that the richer reverberant field spectral content of the Ambisonic Auralisations when compared to typical Binaural processing would indicate that Ambisonics is indeed a superior processing medium with a closer resemblance to natural listening conditions, though comparison to Binaural Impulse Response Auralisation would be required to confirm this.

Timed renders using prototype scripts confirm that the Ambisonic convolution reverb architecture presented by Lopez-Lezanco [13] is possible to apply in real-time using existing FFT and partitioned convolution algorithms.

Time-Frequency Analysis of the developed Switched SIR Ambisonic distance panning function did effectively render B-Format audio dynamically across varying distance. Such renders exhibited the same richer reverberant field and spectral content when compared to typical Binaural processing as was apparent in the static renders. This confirms that the developed system is an appropriate solution to distance-panning in the Ambisonic medium.

Timed renders of dynamic Ambisonic Auralisations using Switched SIR partitioning, FFT and overlap-add algorithms was again suitable for real-time application using existing techniques.

Listening test results were largely inconclusive, though did suggest that the Ambisonic processing more accurately depicts natural listening than Binaural rendering, as was noted several times in a ‘Free Comment’ option given to participants, though this requires further investigation.

The spatial resolution used in this prototype development and analysis was only that of 0.5m for dynamic sound sources. It was noted that little information on Just Noticeable Difference in auditory distance perception is available, and investigation into this would allow for some optimisation in spatial resolution in such systems.

It is also noted that in typical Ambisonic processing rotation of a sound field implies rotation of the acoustic environment. For many applications this is inappropriate, and as such it is proposed that the same functions used in providing Switched SIR Convolution as an Ambisonic Distance Panning function may be extended to include Azimuth and Elevation functions.

## 7. References

- [1] F. Rumsey, Spatial Audio, Oxford: Focal Press, 2001.
- [2] W. Zhang, P. N. Samarasinghe, H. C. and T. D. Abhayapala, “Surround by Sound: A Review of Spatial Audio Recording and Reproduction,” *Applied Sciences*, vol. 7, no. 532, p. 1, 2017.
- [3] E. Benjamin, R. Lee and A. Heller, “Why Ambisonics Does Work,” in *AES 129th Convention*, San Francisco, 2010.
- [4] A. Tarzan, M. Alunno and P. Bientinesi, “Assessment of Sound Spatialisation Algorithms for Sonic Rendering with Headsets,” Aachen university, Aachen, 2017.
- [5] G. Kearney, M. R. H. Gorzel and F. Boland, “Distance Perception in Interactive Virtual Acoustic Environments Using First and Higher Order Ambisonic Sound Fields,” *Acta Acustica United With Acustica*, vol. 98, no. 1, pp. 61-67, 2012.
- [6] A. McKeag and D. McGrath, “Sound Field Format to Binaural Decoder with Head-Tracking,” in *AES 6th Australian Regional Convention*, Melbourne, 1996.
- [7] M. Noisternig, A. Sontacchi, T. Musil and R. Holdrich, “A 3D Ambisonic Based Binaural Sound Reproduction System,” in *AES 24th International Conference on Multichannel Audio*, Banff, 2003.
- [8] T. Nishiyama, T. Nagata, S. Ogata, T. Hasue and M. Kashiwagi, “Recording and mixing techniques for Ambisonic sound reproduction,” in *AES International Conference on Spatial Reproduction*, Tokyo, 2018.

- [9] R. Penha, "Distance Encoding in Ambisonics Using Three Angular Coordinates," in *Sound and Music Computing Conference*, Berlin, 2008.
- [10] D. Menzies and M. Al-Akaidi, "Ambisonic Synthesis of Complex Sources," *Journal of the Audio Engineering Society*, vol. 55, no. 10, pp. 864-875, 2007.
- [11] A. Sontacchi and R. Holdrich, "Distance Coding in 3d Sound Fields," in *AES 21st Conference*, St. Peterburg, 2002.
- [12] B. Wiggins and M. Dring, "AmbiFreeverb 2 - Development of a 3D Ambisonic Reverb with Spatial Warping and Variable Scattering," in *AES Conference on Sound Field Control*, Guildford, 2016.
- [13] Lopez-Lezcano, "An Architecture for Reverberation in High Order Ambisonics," in *AES 137th Convention*, Los Angeles, 2014.
- [14] A. Oliveira, G. Campos, P. Dia, D. Murphy, J. Viera, C. Mendonca and J. Santos, "Real-Time Dynamic Image Source Implementation for Auralisation," in *16th International Conference on Digital Audio Effects*, Maynooth, 2013.
- [15] S. McGovern, "The Image-Source Reverberation Model in an N-Dimensional Space," in *14th International Conference on Digital Audio Effects*, Paris, 2011.
- [16] A. Krokstad, S. Dtrom and S. Sorsdal, "Calculating the Acoustical Room Response by Using the use of a Ray Tracing Technique," *Journal of Sound Vibrations*, vol. 8, no. 1, pp. 118-125, 1968.
- [17] M. A. Gerzon, "Periphony: With Height Sound Reproduction," *Journal of the Audio Engineering Society*, vol. 21, no. 1, pp. 2-10, 1973.
- [18] M. A. Poletti, "Three-Dimensional Surround Sound Systems Based on Spherical Harmonics," *Journal of the Audio Engineering Society*, vol. 53, no. 11, pp. 1004-1025, 2005.
- [19] L. L. Beranek and T. J. Mellow, *Acoustic Sound Fields and Transducers*, London: Academic Press, 2012.
- [20] J. Ahrens, *Analytic Methods of Sound Field Synthesis*, Berlin: Springer, 2012.
- [21] J.-M. Batke, "The B-Format Microphone Revised," in *Ambisonics Symposium*, Graz, 2009.
- [22] C. Nachbar, F. Zotter, E. Deleflie and A. Sontacchi, "Ambix - A Suggested Ambisonic Format," in *Ambisonics Symposium*, Lexington, 2011.
- [23] D. Malham, "Space in Music - Music in Space: Masters Thesis," University of York, York, 2003.
- [24] A. V. Oppenheim and R. W. Schfer, *Discrete Time Signal Processing*, London: Prentice-Hall, 1999.
- [25] G.-B. Stan, J.-J. Ebrechts and D. Archambeau, "Comparison of Different Impulse Response Measurement Techniques," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 249-262, 2002.
- [26] M. Kleiner, B.-I. D. And and P. Svensson, "Auralization - An Overview," *Journal of the Audio Engineering Society*, vol. 41, no. 11, pp. 861-875, 1993.
- [27] J.-M. Jot, L. Cerveau and O. Warusfel, "Analysis and Synthesis of Room Reverberation Based on a Statistical Time-Frequency Model," in *AES 103rd Convention*, New York, 1997.
- [28] N. Tsingos, "Pre-computing geometry-based reverberation effects for games," in *AES 35th International Conference*, London, 2009.
- [29] A. Farina, "Simultaneous Measurement of Impulse Response and Distortion with a Sine Swept technique," in *AES 108th Convention*, Paris, 2000.
- [30] J.-D. Polack and F. L. Figueiredo, "Room Acoustic Auralization with Ambisonics," in *Societe Francaise D'Acoustique*, Nantes, 2012.
- [31] A. Perez-Lopez and J. De Munke, "Ambisonic Directional Room Impulse Response as a new Convention of the Spatially Oriented Format for Acoustics," in *AES 144th Convevntion*, Milan, 2018.
- [32] J. J. Embrechts, "Review on the Application of Directional Impulse Responses in Room Acoustics," in *Congrès français d'acoustique*, Le Mans, 2016.
- [33] U. Zolzer, *Digital Audio Effects*, Chichester: John Wiley and Sons, 2011.
- [34] D. Lyon, "The Discrete Fourier Transform, Part 2: Radix 2 FFT," *Journal of Object Technology*, vol. 8, no. 5, pp. 21-33, 2009.
- [35] F. Wefers, "Partitioned Convolution Algorithms for Real-Time Auralization: PHD Thesis," Logos Verlag, Berlin, 2014.
- [36] R. Stewart and M. Sandler, "Real-time Panning Convolution Reverb," in *AES 12rd Convention*, New York, 2007.
- [37] A. Reilly and D. McGrath, "Real-Time Auralization with Head Tracking," in *AES 5th Australian Regional Convention*, Sydney, 1995.
- [38] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization (Revised Edition)*, London: MIT Press, 1999.
- [39] S. H. Nielsen, "Auditory Distance Perception in Different Room," in *AES 92nd Convention*, Vienna, 1992.



## Full Reviewed Paper at ICSA 2019

Presented\* by VDT.

### The Virtual Acoustic Spaces Unity Spatializer with custom head tracker

T. Resch<sup>1</sup>, M. Hädrich<sup>2</sup>

<sup>1</sup> Hochschule für Musik Basel FHNW, Research and Development | TU Berlin, Audiocommunication Group  
Email: thomas.resch@fhnw.ch

<sup>2</sup> TU Berlin, Audiocommunication Group, Germany, Email: markus.haedrich@tu-berlin.de

#### Abstract

The virtual acoustic spaces (VAS) unity spatializer is a plugin for dynamic binaural synthesis for Unity. It can handle impulse responses (IRs) of arbitrary length (limited only by hardware resources). Hence, it is possible to calculate the binaural synthesis not only with head related transfer functions (HRTFs), but also on the basis of binaural room impulse responses (BRIRs). The plugin can also virtualize reflections calculated by raytracing and it is possible to load an individual IR set for each instance. In addition to being compatible with off-the-shelf cross reality (XR) hardware it features a Bluetooth binding for an easily built, custom-made head tracker based on an ESP32 board. It is therefore predestined for audio augmented reality applications.

#### 1. Introduction

The game engine Unity has become very popular. Not only in the field of game development, but also in scientific areas, for example in the virtualization of acoustic environments. In combination with off-the-shelf XR systems such as the HTC Vive, Unity provides a very simple setup for dynamic binaural synthesis: the angles (azimuth and elevation) between sound source and the person wearing the XR glasses are sent directly from the integrated head tracker to the spatializer plugin. The binaural synthesis is calculated depending on the user's head orientation. But this is only a solution if a complete virtual reality experience is desired. For an audio-only augmented reality (AR), a standalone head tracker is necessary. Furthermore, the existing binaural spatializers for Unity can either not load custom IRs, or at most one set, or the length of the IRs is limited. Most of them are not available for all operating systems. Due to these constraints (which are described in detail in section 2) a first version of the VAS Unity Spatializer was developed for the project Analog Speicher [1] where architecture and the corresponding acoustics of various ancient buildings were simulated. The plugin had to be able to load ten different BRIR sets with lengths of up to 0.5 s into one Unity scene simultaneously. For

the game LosEmal which is currently being developed for the project Myosotis [2], three further requirements were added: firstly, the plugin had to support iOS. Secondly (because the target platform is not a VR system), a connection to a standalone head tracker had to be implemented, because the game principle relies on a well-functioning binaural synthesis. And third, early reflections for less reflecting outdoor environments as described in [3] should be simulated to make the binaural synthesis more plausible.

Therefore, a new version of the plugin was developed with iOS and OSX bindings to an inexpensive, custom-made Bluetooth head tracker, based on an Adafruit Huzzah32 development board. The sensor fusion was realized with a Sparkfun BNO080 inertial measurement unit (IMU) because according to its datasheets, the BNO080 is supposed to perform an outstanding sensor fusion and has not been used in any open source projects yet.

This paper starts with a brief discussion of related work in section 2. Section 3 describes the setup of the native Unity plugin, the corresponding C# scripts, the head tracker and its communication with Unity. Section 4 outlines the implementation details of all components. Section 5 deals with measurement results regarding latency and CPU usage.

## 2. Related Work

Several binaural spatializer plugins are available for Unity. The Oculus plugin [4] supports Android, OSX and Windows. It cannot handle custom IR sets. Microsofts plugin is neither able to do this, nor is it compatible with systems other than Windows [5]. Resonance Audio by Google supports all platforms, but it is not documented how custom HRTFs can be used [6]. Steam Audio provides an SDK and a spatializer for Unity [7]. It supports the Sofa file format [8] and can thus load custom HRTFs and render dynamic binaural synthesis. There is no support for iOS yet and only one IR set can be loaded globally for all plugin instances. The SOFALizer for Unity is capable of loading up to 10 different HRTFs, but the impulse responses are always shortened to 256 samples. According to the developers there is only support for Windows [9]. The Soundscape Renderer (SSR) [10] [11] is a C++ software capable of rendering dynamic binaural synthesis. In combination with a virtual sound device such as Jack it can be used in conjunction with Unity. However, compiling the SSR for iOS or Android is not documented. EVERTims [12] [13] is a framework for the auralization of 3D models with raytracing for OSX, Windows and Linux. It's based on the Accelerated Beam Tracing Algorithm by Lane, Siltanen, Lokki and Savioja [14]. While it looked promising, the Binaural Synthesis Kit [15] is not yet available for download. Open source head tracker projects such as the Hedrot by Alexis Baskind [16], the EDTracker [17], the open headtracker [18] or the MrHeadTracker by Romanov, Berghold, Rudrich, Zauschirm, Frank, Zotter [19] all have a wired transmission only. Robert Twomey's bluetooth-headtracker [20] comes with Bluetooth but the used sensor board is no longer available. The very advanced project DIY-low-cost-head-tracker with sensor fusion, BLE- and serial connection by Sascha Spors [21] uses the MPU9250 which will be deprecated soon.

## 3. Setup and availability

The presented solution consists of four components: the plugin, the scripts for plugin configuration, the head tracker and a small Bluetooth app with two corresponding Unity scripts which enable the communication between head tracker and Unity. In its simplest configuration, the plugin renders a dynamic binaural synthesis with the possibility to apply a directional pattern to the sound emitter. In order to take full advantage of the Unity environment, the plugin can be configured to calculate up to 20 reflections. All components are available as source code at the project repository [22] including precompiled plugin binaries for iOS and OSX, a sample scene for Unity and detailed installation instructions. Head tracker firmware, circuit diagram and additional information for building, programming and configuration are also available there. Detailed calibration instructions for the BNO080 are provided by the manufacturer [23].

### 3.1. Unity

The plugin binary must be placed in the Unity project folder in `Assets/Plugins/(TARGET_PLATFORM)` and the

VAS\_Unity\_Spatializer must be chosen as *Spatializer Plugin* under *Project Settings/Audio*. Unity will search automatically for the version appropriate for the respective target platform. IRs must be placed within the *Assets/StreamingAssets* folder to ensure cross-platform file access. In the settings of any used audio source the checkbox *Spatialize* must be activated and *Spatialize Blend* should (usually) be set to 1. In order to load an IR set into a plugin instance, one of the C# *VasSpatConfig* scripts must be added to the Unity Game Object that contains the audio source. Three different versions are available:

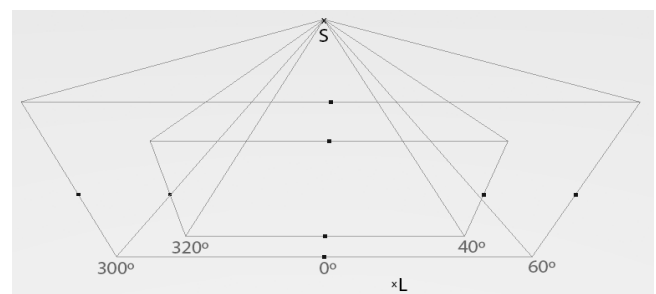
- VasSpatConfigSimple
- VasSpatConfigManual
- VasSpatConfigAuto

All three implement the basic communication between the native audio plugin and C#. The latter one demonstrates the usage of Unity's physics engine in cooperation with the VAS Unity Spatializer for raytracing.

#### 3.1.1. VasSpatConfigSimple

This script configures the plugin as a simple binaural renderer. Audio sources can be provided with a directional pattern. The script exposes seven variables in the Unity editor view:

- *IR set* has to be set to the IR filename including its extension but without its path. Supported file types are .txt files in the VAS format and Sofa files [8].
- *Global* denotes whether the IR set should be used as a global filter for all instances of the plugin.
- *Directivity damping* defines, whether the signal is damped linearly or logarithmically outside its full sound pressure area.
- *Horizontal source width* sets the area in degrees in the horizontal plane where the source is audible.
- *Horizontal full sound pressure* defines the area where the signal is emitted with full sound pressure in the horizontal plane.
- *Vertical source width* sets the directivity in degrees in the vertical plane where the source is audible.
- *Vertical full sound pressure* defines the area where the signal is emitted with full sound pressure in the vertical plane.



**Fig. 1:** Directional pattern example with a horizontal source width of 120° and full sound pressure level of 80°.

If the four latter parameters are set to 360°, the source behaves as an omnidirectional emitter. If the horizontal width parameter is, for example, set to 120° and the full horizontal



sound pressure to  $80^\circ$  (fig 1.), the source emits from  $0^\circ$  to  $40^\circ$  and from  $320^\circ$  to  $0^\circ$  with full energy. Within  $40^\circ$  to  $60^\circ$  and  $330^\circ$  to  $340^\circ$  the signal is gradually lowpass-filtered and attenuated. This is done either linearly or logarithmically (depending on the Directivity damping parameter). Bi-directional patterns can be achieved with two sources, emitting into opposite directions. Distance related damping is realized with Unity's build-in audio source features.

### 3.1.2. VasSpatConfigManual

With the VasSpatConfigManual script it is possible to add five binaural reflections for an Audio Source. The user has to manually create and place game objects in the Unity scene and drag them onto the public variable slots of the script. They determine the locations of the corresponding reflections. Public variables in addition to those of the first script are:

- *Reflection 1–5* are public variable slots for arbitrary game objects representing the position of the reflections.
- *Material stiffness* selects a material characteristic. Possible settings are low, middle and high.

### 3.1.3. VasSpatConfigAuto

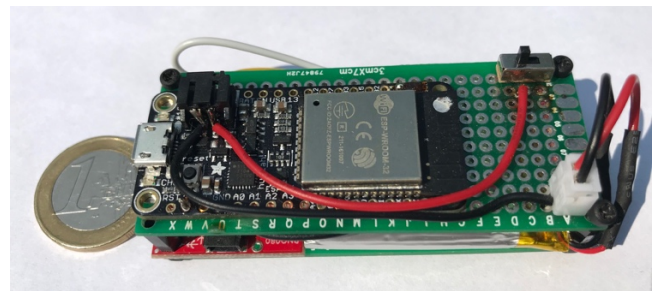
This script uses raytracing to determine the locations of the reflections which are updated in real time. The number of rays is currently hardcoded to five rays. It has one additional variable:

- *Reflection order* defines, how many reflections per ray shall be calculated.

## 3.2. Head tracker

Precision and latency of the head tracking are among the most important quality assurance factors for an immersive auralization of virtual acoustic scenes, eg. dynamic binaural synthesis. For the entire audio path, refresh rates of 60 Hz and total delay times of 50 ms are considered acceptable [24]. Because head tracking is only the first part of this audio path, less latency at this point leaves more time for subsequent audio processing and is therefore desirable. The practical aim of the presented solution is an interaction in which both source and listener are moving. Therefore, the minimal audible movement angles (MAMA) [25] are relevant factors. Strybel, Manligas and Perrott found a sensitive area for movement detection with  $1^\circ$  to  $2^\circ$  at a position of  $+40^\circ$  and  $-40^\circ$  azimuth and elevations below  $80^\circ$ . Outside of this area the MAMA increased to  $3^\circ$  to  $10^\circ$  [26]. Keeping this in mind the head tracker device should offer a minimum accuracy of less than  $3^\circ$ . The presented do-it-yourself (DIY) low cost head tracker device is made of an Host MCU - Adafruit Huzzah32 development board (ESP32), which supports Wifi and Bluetooth Low Energy (BLE) 4.2, and a Sparkfun BNO080 IMU sensor board connected via I2C. Apart from receiving the IMU data, the ESP32 handles the wireless communication and device management. For mobile use, the device has its own power supply, in the form of a 3.7 V LiPo battery, and a hardware on/off switch. The BNO080 provides orientation data with and without inclusion of the magnetometer. The internal sensor fusion uses the magnetometer for drift correction of

the gyroscope. Thus, a smooth output (e. g. for games) or the most accurate output can be selected. The former setup can lead to the typical drift in long-term applications, the second setup to possible jumps during the correction process. However, with the help of a stabilization function (AR/VR stabilization), these jumps can be gradually corrected so that this sensor board is well suited for AR/VR tracking applications. When using the *Gaming Rotation Vector*, which does not use the magnetometer, a static/dynamic error of  $1.5^\circ/2.5^\circ$  is specified. This complies with the Strybel et. al. [26] condition for the MAMA. In this setup, the drift of  $0.5^\circ/\text{min}$  can be balanced if AR/VR stabilization of this vector is selected [27].



**Fig. 2:** VAS head tracker consisting of an Adafruit Huzzah32 development board, a Sparkfun BNO080 IMU sensor board and a mobile power supply, (LiPo battery, 3.7 V).

Azimuth and elevation do not describe the exact head position of the listener, as a possible lateral tilt of the head is not included. There is currently no HRTF or BRIR dataset that also shows lateral tilt of the head on a straight torso. The advanced Head-Above-Torso (HATO) HRTF database created by Brinkmann et. al. [28] also uses only azimuth and elevation. In the future, if there are data sets that support a lateral head tilt (Euler angle: roll), this angle can easily be provided by the BNO080. However, such data sets would either be very large, since for every possible head tilt a complete  $360^\circ$  data set would have to be present or would have very high computational costs due to the interpolation required for reduced data sets. Both cases are rather unfavorable in terms of resource allocation for mobile use and require further development work, both hardware and software.

Since both, the BNO080 and Unity work internally with quaternions with the y-axis (here elevation) is limited to  $\pm 90^\circ$  [27, 29], no problems with the gimbal-lock are known.

### 3.3. Head tracker connection to Unity

VAS Head Tracker Connect is a standalone software, currently available for iOS and OSX, that serves as an intermediary between Unity and head tracker. It connects to the head tracker via Bluetooth and sends azimuth and elevation as open sound control (OSC) UDP packets to Unity. Two values can be set in the user interface:

- *OSC port number* must be set to match the receiver port in Unity
- *Headtracker ID* must be set to match the head tracker's name which is currently hardcoded to the head trackers firmware.



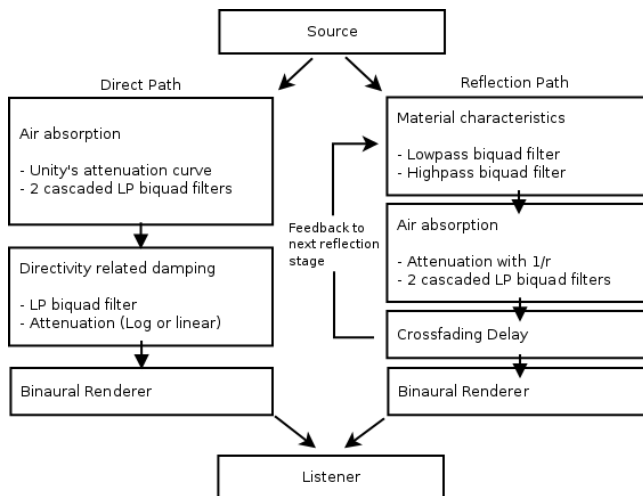
In Unity the package uOSC [30] has to be installed. The scripts uOSCServer and ReceiveHeadtrackerData have to be attached to the *Audio Listener* object.

## 4. Implementation details

The plugin performs a uniformly partitioned overlap add convolution. Length of the IRs is not limited (only by hardware resources). For implementation details about the underlying rendering engine, please refer to the publication and documentation about the VAS library [22] [31].

### 4.1. Unity

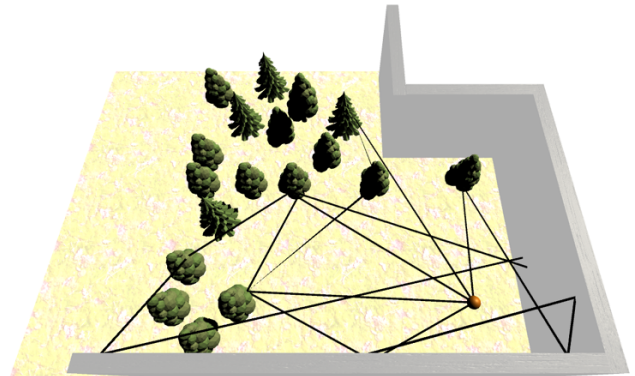
The plugin is implemented with Unity’s native audio plugin SDK in C++. Azimuth and elevation are automatically accessible in C++ for the direct path from source to listener. To be able to calculate angles and delay times for the reflections, their positions must be transmitted manually as float values from C#, along with the parameters for material characteristics. The maximum number of reflections is limited by the native plugin parameters which have to be declared and initialized in advance in the data structure of the plugin. The current implementation uses five rays. Reflections are calculated up to the 4th order, resulting in twenty reflections in total. The complete signal processing pipeline is illustrated in figure 3 below.



**Fig. 3:** The complete signal processing pipeline from source to listener.

Frequency related air absorption is approximated for 20° C and 20 % humidity with two cascaded biquad filters, similar to the illustrated filter curves in [32]. The directional pattern (and the corresponding damping) is applied to the direct path from sound emitter to receiver only. Reflections are considered to be omnidirectional. Their positions are calculated with the basic raytracing technique as described in [33]. As recommended by the authors, a predefined distribution pattern is used due to the small number of rays. Schröder suggests distributing them evenly across the source’s surface [34]. In the presented solution, sound sources are considered as points without volume. Therefore, rays are evenly distributed within the source’s directional pattern in

the horizontal plane as shown in figure 4. Material characteristics are currently modelled in a simplified manner using a lowpass and a highpass biquad filter.



**Fig. 4:** Five evenly distributed rays with a directional width of 120° in a Unity scene with 4th order reflections.

Since outdoor environments usually have no ceiling, reflections of a higher order would no longer reach the listener if elevation angles are too large. Therefore, they are randomly varied by  $\pm 2^\circ$  only. Only specular reflections are calculated. The possibility of diffuse reflections is currently ignored. The delay line is realized with two delays. In the moment the delay time changes, both the current and the target delay are performed and a crossfade with a length of 1024 samples is calculated from current to target delay. This makes large jumps possible without artefacts and prevents the typical pitch shifting effects of interpolated delays.

### 4.2. Head tracker

Overall latency for head tracking consists of the packet delivery time  $\tau_{total}$ , which is the sum of packet transmission time  $\tau_{trans}$  and the sensor device latency  $\tau_{IMU}$ . With a polling time  $\tau_{poll}$  at a sample rate of 200 Hz and a propagation delay  $\tau_{prop}$  of 3.7 ms [23],  $\tau_{IMU}$  is about 8.7 ms. The transmission time  $\tau_{trans}$  is determined by the Bluetooth LE (BLE) transmission speed of the connection between the ESP32 and the mobile device. BLE uses channel hopping and so its communication consists of a consecutive number of connection events, organized at a specific connection interval  $\tau_{ci}$ . After one  $\tau_{ci}$  the frequency channel will be switched. The connection parameters are initially determined when a connection is established.

In the presented soft- and hardware solution the computer or smartphone acts as the master and finally defines  $\tau_{ci}$  to the peripheral slave head tracker device. However, the peripheral device may ask for certain connection parameters. In case the suggested parameters do not meet the specifications of the central device, the request will be rejected. Depending on the operating system and the device generation, this minimum connection interval ranges from 7.5 ms to 30 ms and the maximum number of packets per connection interval  $N_{ci}$  may be 4, 6 or 7. Because BLE is a shared resource on mobile devices, the operating system can scale down  $N_{ci}$  as needed and increase  $\tau_{ci}$  as needed. In central mode the connection parameters are determined by iOS. On the iPhone 8 test device, a  $\tau_{ci}$  of 15 ms [35] and a  $N_{ci}$  of 7 is supported.

The data (azimuth and elevation as CSV) takes 8 bytes per orientation event, which fits easily into the default maximum transfer unit (MTU) size of 23 bytes which has a possible payload of 20 bytes, so only one packet per orientation event is needed. In line with our requirement of transmitting small data sizes within a strict time limit we need the smallest possible effective connection interval  $\tau_{eci}$  and the highest  $N_{ci}$ . Therefore, the slave latency, the number of skipped connection intervals,  $N_{sl}$  is set to zero with  $\tau_{eci} = \tau_{ci}$  [36] and the peripheral will send an update request to the central to ask for the smallest  $\tau_{ci}$ .

## 5. Results

### 5.1. Latency

The predicted latency  $\tau_{total}$  of the head tracking device in conjunction with the iPhone 8 test device should be:

$$\tau_{total} = \tau_{trans} + \tau_{IMU}$$

$$\tau_{trans} = \frac{\tau_{eci}}{N_{pack}} \quad \tau_{eci} = \tau_{ci}(1 + N_{sl})$$

$$\tau_{IMU} = \tau_{poll} + \tau_{prop}$$

$$\tau_{total} = \frac{15 \text{ ms}}{7 \text{ packets}} + 5 \text{ ms} + 3.7 \text{ ms} = \frac{10.8 \text{ ms}}{\text{packet}}$$

with  $\tau_{ce} < \frac{\tau_{eci}}{N_{pack}}$  to be able to use the maximum allowed number of packets.

With the capability of BLE to transmit 1 symbol in 1  $\mu\text{s}$  [37], the time for a single connection event consisting of one communication cycle can be estimated as follows [36]:

1. Receive packet with a payload of 8 bytes and a maximum protocol overhead of 14 bytes [37] ( $22 \times 8 \text{ bit} \times 1 \mu\text{s}$ ),
2. Mandatory interframe space (150  $\mu\text{s}$ ),
3. Send acknowledge packet ( $80 \text{ bit} \times 1 \mu\text{s}$ ),
4. Mandatory interframe space (150  $\mu\text{s}$ ).

The sum meets the above condition for  $\tau_{ce}$ :  $0.536 \text{ ms} < 2.14 \text{ ms}$ .

A theoretical number of possible  $N_{pack}$  per  $\tau_{ci}$  can be calculated with:

$$N_{pack} = \frac{\tau_{ci}}{\tau_{ce}}$$

Under real world conditions the influence of the bit error rate (BER), as demonstrated by Gomez, Demirkol and Paradells [38], the interference with other devices using the 2.4 GHz band, packet loss [36] and the restrictions of the central operating system leads to a much lower transmission rate and varying latencies. Especially the limitation of  $N_{pack}$  per  $\tau_{ci}$  by the operating system means that no further data

exchange takes place after reaching the maximum  $N_{pack}$  until the end of the  $\tau_{ci}$ .

For the measurement 10,000 numbered packets were sent in ten iterations from the peripheral to the central device at a distance of 1 m and with a received signal strength indicator (RSSI) of  $-70 \pm 5 \text{ dB}$ . By using offline logging and a subsequent evaluation of timestamps and quantity of sent and received packets, an average packet loss of  $< 1.5 \%$  was measured.

Packet loss leads to sporadically occurring higher latencies, which directly affects the update rates of the central device. Therefore, only an averaged update rate of appr. 5 ms at a IMU refresh rate of 200 Hz can be considered. With this setup, we estimate an average latency of the head tracking system of appr. 11 ms.

Using the *Best Latency* setting in Unity's audio preferences leads to a vector size of 256 samples under both operating systems (iOS in conjunction with an iPhone 8, and OSX) which corresponds to 5.8 ms (assuming a sample rate of 44.1 kHz). The dynamic filter change and the resulting crossfade between current and new angle causes an additional latency of 11.6 ms. The average OSC transmission time from the VAS Head Tracker Connect software to Unity was measured on a Macbook Pro 2018 (14.4 ms) and an iPhone 8 (18 ms).

At a refresh rate of 200 Hz this results in a total system latency of 42.2 ms on OSX and 46.2 ms on iOS, which meets the above-mentioned criterion [24]. The default latency setting, which leads to a vector size of 1024 samples on iOS, leads to a total latency of 63.6 ms. This value could still be considered acceptable, but savings in terms of CPU load are almost negligible (see table 1).

### 5.2. CPU load

CPU load was measured with Xcode Instruments on an iPhone 8. Partition and FFT size for the convolution were set to match the vector size. The percentage value in the right column is the CPU load for one core for one voice.

Vector size	CPU load (iPhone 8, one core)
256	9 %
1024	8 %

Tab. 1: CPU load on an iPhone 8.

A voice includes the playback of the audio source, Unity internal DSP (distance attenuation, doppler effect, mixer) and the complete signal processing of the plugin with 20 reflections. The head tracker was turned constantly, so that the convolution for the binaural synthesis (with an HRTF length of 256 samples) had to be carried out continuously for the current and the target angle.

## 6. Conclusion and outlook

The presented soft- and hardware is a powerful and easily configurable engine for rendering dynamic binaural synthesis in Unity. Besides real time calculation of HRTF based

binaural synthesis including up to 20 reflections it can process BRIRs of arbitrary length (only limited by hardware resources). The possibility to load an individual IR set for every plugin instance makes the VAS Unity Spatializer unique for the time being. This enables the user to, for example, equip different rooms with different BRIRs, preload several IR sets for listening tests or allow multiple users to experience one scene with different (for instance individualized) HRTFs simultaneously.

In Unity, the audio vector size is not as finely adjustable as in other environments, especially those focused primarily on audio (such as Pure Data or Max/MSP) where sizes as small as 16 samples are achievable. However, due to the low latency of the presented head tracker, the overall system latency is well within the requirements for dynamic binaural synthesis.

The Adafruit board can be configured as a Wifi access point. With the next firmware version, it will be possible to set all parameters (data format: e. g. euler angles or quaternions, connection type, sensor fusion method) via a static page hosted on the board. In order to enable use in environments without Bluetooth, data transmission via Wifi and OSC will be implemented. For the presented raytracing solution, physical principles have been simplified to ensure good usability and not to overuse hardware resources on the iOS platform. The focus was on outdoor environments with little reflections. A future release will enable the user to use more natural directional patterns, different material characteristics and a much larger number of reflections.

## 7. References

- [1] HZK, "Auralisierung archäologischer Räume", [Online]. Available: <https://www.interdisciplinary-laboratory.hu-berlin.de/de/content/analogspeicher-ii-auralisierung-archaologischer-raume>. [Accessed 29. 08. 2019].
- [2] FHNW, "FHNW Mysotis Garden," [Online]. Available: <https://www.fhnw.ch/de/die-fhnw-hochschulen/ht/institute/institut-fuer-data-science/fhnw-myosotis-garden>. [Accessed 22. 06. 2019].
- [3] F. Stevens, D. T. Murphy, L. Savioja and V. Välimäki, "Modeling Sparsely Reflecting Outdoor Acoustic Scenes Using the Waveguide Web," *IEEE/ACM Transactions On Audio, Speech, and Language Processing*, p. pp. 1566–1578, 08. 2017.
- [4] Oculus, "Oculus Spatializer," [Online]. Available: <https://developer.oculus.com/downloads/package/oculus-spatializer-unity/>. [Accessed 21. 06. 2019].
- [5] Microsoft, "Microsoft Mixed Reality Documentation," [Online]. Available: <https://docs.microsoft.com/en-us/windows/mixed-reality/spatial-sound-in-unity>. [Accessed 21. 06. 2019].
- [6] Google, "Resonance Audio," [Online]. Available: <https://resonance-audio.github.io/resonance-audio/>. [Accessed 21. 06. 2019].
- [7] Steam Audio, "Git Repository Steam Audio," [Online]. Available: <https://valvesoftware.github.io/steam-audio/downloads.html>. [Accessed 27. 02. 2019].
- [8] P. Majdak, Y. Iwaya, T. Carpentier, R. Nicol, M. Parmentier, A. Roginska, Y. Suzuki, K. Watanabe, H. Wierstorf, H. Ziegelwanger und M. Noisternig, „Spatially Oriented Format for Acoustics: A Data Exchange Format Representing Head-Related Transfer Functions,“ in *Proceedings of the 134th Convention of the Audio Engineering Society*, Rom, 2013.
- [9] M. P. Jenny C. and C. Reuter, "SOFA Native Spatializer Plugin for Unity - Exchangeable HRTFs in Virtual Reality," in *Proceedings of the 144th Convention of the Audio Engineering Society*, Milan, 2018.
- [10] M. Geier, J. Ahrens and S. Spors, "SoundScape Renderer," [Online]. Available: <http://spatialaudio.net/ssr/>. [Accessed 10. 02. 2019].
- [11] M. Geier, J. Ahrens und S. Spors, „The SoundScape Renderer, A unified spatial audio reproduction framework for arbitrary rendering methods,“ in *124th AES Convention*, Amsterdam, 2008.
- [12] LIMSI/CNRS, TKK/Department of Media Technology, IRCAM, "EVERTims," [Online]. Available: <https://evertims.github.io>. [Accessed 27. 02. 2019].
- [13] M. Noisternig, B. Katz, S. Siltanen and L. Savioja, "Framework for real-time auralization in architectural acoustics," *Acta Acustica United with Acustica*, vol. 94, no. 6, p. 1000–1015, 2008.
- [14] S. Laine, S. Siltanen, T. Lokki und L. Savioja, „Accelerated beam tracing algorithm,“ *Applied Acoustics*, Bd. 70, Nr. 1, p. 172–181, 2009.
- [15] A. Franck, G. Costantini, C. Pike and F. M. Fazi, "An Open Realtime Binaural Synthesis Toolkit for Audio Research," in *Audio Eng. Soc. 144th Conv*, Milano, 2018.
- [16] A. Baskind, "Hedrot," [Online]. Available: <https://abaskind.github.io/hedrot/>. [Accessed 21. 06. 2019].
- [17] V. Manoukian, "EDTracker2," [Online]. Available: <http://www.edtracker.org.uk>. [Accessed 21. 06. 2019].
- [18] D. Frie, "DIY Headtracker (Easy build, No drift, OpenSource)," [Online]. Available:

- <http://www.rcgroups.com/forums/showthread.php?t=1677559>. [Accessed 21. 06. 2019].
- [19] M. Romanov, P. Berghold, D. Rudrich, M. Zaunschirm, M. Frank and F. Zotter, "Implementation and Evaluation of a Low-cost Head-tracker for Binaural Synthesis.," in *142th AES Convention*, Berlin, 2017.
- [20] R. Twomey, "bluetooth-headtracker," [Online]. Available: <https://github.com/roberttwomey/bluetooth-headtracker/tree/d3df1d65b69e2e189bb189d9948c26a76d16ca1a>. [Accessed 21. 06. 2019].
- [21] S. Spors, "diy-low-cost-head-tracker-2," [Online]. Available: <http://spatialaudio.net/diy-low-cost-head-tracker-2/>. [Accessed 21. 06. 2019].
- [22] T. Resch, „Git Repository VAS Library,“ [Online]. Available: [https://github.com/funkerresch/vas\\_library](https://github.com/funkerresch/vas_library). [Zugriff am 22 06 2019].
- [23] Hillcrest Labs, "BNO080/BNO085 Sensor Calibration Procedure," [Online]. Available: <https://www.hillcrestlabs.com/downloads/bno080-sensor-calibration-procedure>. [Accessed 17. 06. 2019].
- [24] M. Vorländer, Auralization, Heidelberg: Springer Berlin, 2008.
- [25] D. W. Chandler and D. Grantham, "Minimum audible movement angle in the horizontal plane as a function of stimulus frequency and bandwidth, source azimuth, and velocity.," *J. Acoust. Soc. Am.*, Vol. 91, No. 3., p. pp. 1624–1636, 03. 03. 1992.
- [26] T. Strybel, C. L. Manligas and P. D. R. ., "Minimum audible movement angle as a function of the azimuth and elevation of the source," *Human Factors The Journal of the Human Factors and Ergonomics Society*, p. 267–275, 07. 1992.
- [27] Hillcrest Labs, "BNO080/BNO085 Datasheet," [Online]. Available: <https://www.hillcrestlabs.com/downloads/bno080-datasheet>. [Accessed 17. 06. 2019].
- [28] F. Brinkmann, A. Lindau, S. Weinzierl, S. v. d. Par, M. Müller-Trapet, R. Opdam and M. Vorländer, "A High Resolution and Full-Spherical Head-Related Transfer Function Database for Different Head-Above-Torso Orientations," *J. Audio Eng. Soc.*, vol. 65, no. 10, p. 841–848, 2017.
- [29] Unity Technologies, "Unity Documentation," [Online]. Available: <https://docs.unity3d.com/Manual/QuaternionAndEulerRotationsInUnity.html>. [Accessed 20. 08. 2019].
- [30] Hecomi, „Git Repository uOSC,“ [Online]. Available: <https://github.com/hecomi/uOSC>. [Zugriff am 22 06 2019].
- [31] C. B. S. W. T. Resch, „VAS – A cross platform C-library for efficient dynamic binaural synthesis on mobile devices,“ in *AES, International Conference on Headphone Technology*, San Francisco, 2019.
- [32] L. S. M. K. J. Huopaniemi, „Modeling of reflections and air absorption in acoustical spaces — A digital filter desing,“ in *Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, USA, 1997.
- [33] U. P. S. L. Savioja, „Overview of geometrical room acoustic modeling techniques,“ *The Journal of the Acoustical Society of America* 138, p. 708–730, 2015.
- [34] D. Schröder, „PHYSICALLY BASED REAL-TIME AURALIZATION OF INTERACTIVE VIRTUAL ENVIRONMENTS,“ *Aachener Beiträge zur technischen Akustik 11*, 2011.
- [35] Apple, "Technical Q&A QA1931 Using the correct Bluetooth LE Advertising and Connection Parameters for a stable connection," [Online]. Available: [https://developer.apple.com/library/archive/qa/qa1931/\\_index.html](https://developer.apple.com/library/archive/qa/qa1931/_index.html). [Accessed 05. 06. 2019].
- [36] J. Afonso, A. Maio and R. Simoes, "Performance Evaluation of Bluetooth Low Energy for High Data Rate Body Area Networks," *Wireless Personal Communications*, p. 121–141, 09. 2016.
- [37] Bluetooth SIG, Inc., "CS – Core Specification," [Online]. Available: [https://www.bluetooth.org/docman/handlers/download.doc.ashx?doc\\_id=441541](https://www.bluetooth.org/docman/handlers/download.doc.ashx?doc_id=441541). [Accessed 23. 06. 2019].
- [38] C. Gomez, I. Demirkol and J. Paradells, "Modeling the Maximum Throughput of Bluetooth Low Energy in an Error-Prone Link," *IEEE COMMUNICATIONS LETTERS*, vol. 15, no. 11, p. 1187–1190, 11. 2011.





# Full Reviewed Paper at ICSA 2019

Presented by VDT.

## Room geometry inference using sources and receivers on a uniform linear array

Youssef El Baba<sup>1</sup>, Andreas Walther<sup>2</sup>, Emanuel A. P. Habets<sup>3</sup>

<sup>1</sup> *International Audio Laboratories Erlangen<sup>†</sup>, Germany, Email: youssef.elbaba@audiolabs-erlangen.de*

<sup>2</sup> *Fraunhofer Institute for Integrated Circuits, Erlangen, Germany, Email: andreas.walther@iis.fraunhofer.de*

<sup>3</sup> *International Audio Laboratories Erlangen<sup>†</sup>, Germany, Email: emanuel.habets@audiolabs-erlangen.de*

### Abstract

State-of-the-art room geometry inference algorithms estimate the shape of a room by analyzing peaks in room impulse responses. These algorithms typically require the position of the source wrt the receiver array; this position is often estimated with sound source localization, which is susceptible to high errors under common sampling frequencies. This paper proposes a new approach, namely using an array with a known geometry and consisting of both sources and receivers. When these transducers constitute a uniform linear array, new challenges and opportunities arise for performing room geometry inference. We propose solutions designed to address these challenges, but also designed to leverage the opportunities for better results.

Keywords: Image model, time of arrival disambiguation, echo labeling, reflection point localization, reflector localization, room geometry inference.

## 1. Introduction

The task of room geometry inference (RGI) is concerned with the localization of reflective boundaries in an enclosed space, and is of interest in several applications [1]: 3D sound analysis and reproduction, robust sound source localization (SSL), speaker tracking and de-reverberation. RGI methods use times of arrival (TOAs) of the direct-path and reflections — peaks in room impulse responses (RIRs) from different microphone and loudspeaker position combinations — to infer the locations and orientations of planar reflectors. In specific, first-order TOAs characterize the physical walls present in the room. The largest family of reflector localization (RL) methods relies on ellipse geometry [2–6] or hyperbola geometry [7–9]. Other methods rely on beamforming or other schemes [1, 10]. For RL, TOAs need to be separated into

sets, each set belonging to a single reflector [8]. These sets are used individually with the measurement position, either known or estimated using SSL, to define multiple constraints which together localize a reflector.

RGI can considerably benefit from a-priori knowledge of all the transducers' locations. Most importantly, finding the system latency in real measurements is a challenge [1] which can be alleviated with knowledge of the relative transducer positions. Known array geometries are commonly assumed in the RGI literature [1]; however, these usually contain either microphones or loudspeakers, exclusively. Employing an array with both types of transducers is uncommon<sup>1</sup>. Nonetheless, existing arrays with a single type of transducer can be transformed into arrays having both types by using one loudspeaker as a microphone or vice versa; this is made possible by acoustic transducer reciprocity. Thus, a known array

<sup>†</sup>A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer IIS, Germany.

<sup>1</sup>Albeit there are exceptions setting a precedent for this [11].



geometry can be equivalent to known relative loudspeaker-microphone positions; this motivates our adoption of an intra-array RGI setup involving a uniform transducer array with multiple loudspeakers and one microphone (non-coincident).

This paper presents multiple adaptations to our existing RGI algorithm [1] to address the challenges of this intra-array setup, e.g., those due to shorter distances between sources and the receivers. Additionally, the paper proposes one new improvement inspired by this intra-array setup and leveraging the opportunities offered by it, as well as two more general improvements independent of the setup. The novelties and performance evaluation are presented in 2D; however, they are generalizable to 3D.

## 2. Room geometry inference problem and existing solution

### 2.1. Problem formulation

Given a uniform linear array (ULA) of  $L$  loudspeakers with a single omnidirectional microphone, and assuming that the acoustic propagation can be modeled by a linear time-invariant filter<sup>2</sup>, the RIR of the filter between the  $j$ -th loudspeaker and the microphone (notwithstanding noise) can be expressed by

$$h_j(t) = \alpha_{0j} \delta(t - \tau_{0j}) + \sum_{r=1}^R \alpha_{rj} \delta(t - \tau_{rj}), \quad (1)$$

where  $\alpha_{0j}$  and  $\alpha_{rj}$  are the attenuation coefficients of the direct and reflection paths, respectively, the index  $r$  refers to one of  $R$  real or image reflectors, the function  $\delta(t)$  represents the delta function and  $t$  denotes time. The TOAs  $\tau_{0j}$  that arrive from the  $L$  loudspeakers to the real microphone and the TOAs  $\tau_{rj}$  ( $r \in \{1..R\}$ ) that arrive to the image microphones correspond to the direct and reflected wavefronts, respectively; they form the sets  $\mathcal{T}_r = \{\tau_{rj} : \forall j \in \{1..L\}\}$ .

These RIRs and the known relative positions of the loudspeakers and microphone in the array constitute the input data. TOAs need to be detected and disambiguated into separate sets  $\{\mathcal{T}_r : \forall r \in \{0..R\}\}$ . The aim is to obtain from these TOA sets the desired plane equations  $\langle \mathbf{n}_r, \mathbf{x} \rangle + o_r = 0$  characterizing the different reflectors' planes<sup>3</sup>, where  $\langle \cdot, \cdot \rangle$  denotes the scalar product between vectors,  $\mathbf{n}_r$  and  $o_r$  denote the  $r$ -th plane's normal vector and offset, and  $\mathbf{x}$  denotes the Cartesian 2D coordinate vector.

<sup>2</sup>Although RIRs simulated with this image-source model [12] differ from those measured in reality, namely due to model errors, the model reproduces the early wavefronts' arrival times with sufficient accuracy for our application. This is because RGI only uses early (first- or at most second-order) reflections in rectangular rooms, and the wavefronts these produce are negligibly affected by inaccuracies of the model in simulating modal behavior or taking into account frequency-dependent absorption etc.

<sup>3</sup>Thus, finite reflectors are approximated by infinite planes. The final, finite room geometry can be obtained after the algorithm selects the planes corresponding to physical walls present in the room (after the region-spot-searching mode described in Section 3.2): these infinite planes intersect precisely at the boundaries of the physical walls.

### 2.2. Overview of existing solution

We build upon our existing RGI method from [1], but we do not require the graph-based 3D extension it includes. This method consists of four steps. First, peaks corresponding to TOAs in the RIRs are detected and labeled using the linear Radon transform (LRT) [13]. Second, the labeled TOA sets are used to estimate the image microphone positions using [14], with knowledge of the source-receiver array geometry. Third, using the estimated image microphone positions and the array geometry, the positions of reflection points on the available reflectors are determined using the RL method in [5]. Finally, the reflection points determine the reflectors' locations and orientations. In addition to the known array geometry, this method assumes a known speed of sound and sampling frequency, which is equal across all transducers. In the case of real measurements, it also assumes zero inter-transducer latency, while allowing for a known global latency.

### 2.3. Challenges with intra-array setup

In this work, we assume the ULA is placed near and parallel to a reflector in the room. We use one loudspeaker in the array as a microphone; other loudspeakers are operated normally, not reciprocally. The main challenge in this setup is the near-field scenario due to the short distances between sources and receivers. This is only mitigated with lower sampling frequencies, which have the negative side effect of decreasing the precision of the LRT.

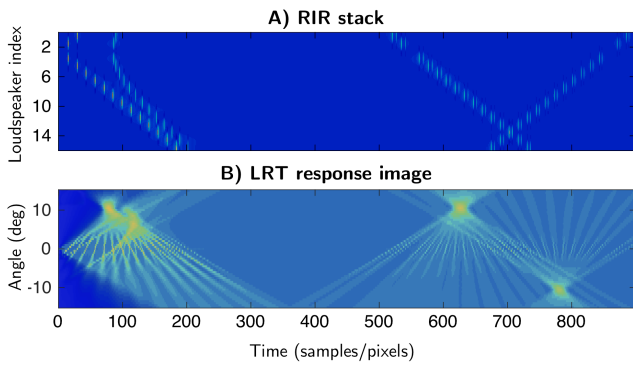
The direct sound from the nearest loudspeaker to the microphone arrives shortly after  $t = 0$ , and is thus disproportionately louder than the sound arriving in reflections or from farther loudspeakers; this is due to the  $1/r$  sound attenuation law: in the near-field region (small distances  $r$ ), differences in attenuation can be drastic between sound paths of different lengths. This causes the RIRs corresponding to the loudspeakers near the microphone to be dominated by their direct sound peaks, with reflections in these RIRs or even direct sound from other RIRs becoming relatively negligible; this translates in turn into disproportionately faint reflection responses on the LRT, especially for microphones positioned centrally on the ULA.

On the other hand, the near-field scenario violates the far-field assumption in the LRT [13]; this problem is especially noticeable for the direct sound and the wavefront reflected from the nearest wall (see Fig. 1). These wavefronts can no longer be accurately considered planar as they exhibit high curvature: the main lobes of their LRT responses are accordingly more diffuse, spread over a bigger temporal/angular region on the LRT, they attain a lower maximum amplitude and are splintered into multiple sub-responses.

## 3. Proposed adaptations and improvements to existing solution

### 3.1. Near-field adaptations

A significant contribution of this work are four adaptations designed to counter the artifacts of working in a near-field



**Fig. 1:** Example resampled RIR stack (A) and its LRT response image (B) for a near-field scenario (Setup 1, microphone position 3 in Section 4.1, the stack is re-attenuated (see Section 3.1), and both the stack and the LRT are enhanced here for visualization). Yellow encodes high values, blue encodes low values. The two figures share the same horizontal (time) axis but have different vertical axes. Notice the lower focus of the main lobes of the LRT responses for the two earliest near-field wavefronts (in upper left region in (B)), with respect to the main lobes of the LRT responses for the later wavefronts (around samples 600-800 in (B)).

scenario.

The first adaptation selectively re-attenuates the disproportionally-boosted direct sound and earliest reflections wrt the later reflections in the RIRs, both within and outside wavefronts. Only the early region is attenuated as it is not desirable to simply compensate for the  $1/r$  law for all samples: this would significantly increase noise levels in the later portion of the RIRs, with deleterious effects for the LRT; moreover, extending the re-attenuation region to later portions is also of little use since later reflected wavefronts do not suffer from near-field effects. The procedure first computes the reference direct sound TOAs  $T_{j,ref}$  for all loudspeakers  $j = 1..L$  using the array geometry, then detects the earliest TOAs in each actual RIR using a peak picker; it then compares these two TOA sets to estimate any inherent global latency in the RIRs<sup>4</sup>. Within a temporal neighborhood  $T_{er}$  around the (latency-corrected) direct sound peaks, we re-attenuate<sup>5</sup> the RIRs via multiplication by  $r$  (translated into time) according to  $h'_j(t) = h_j(t)f_j(t)$  with

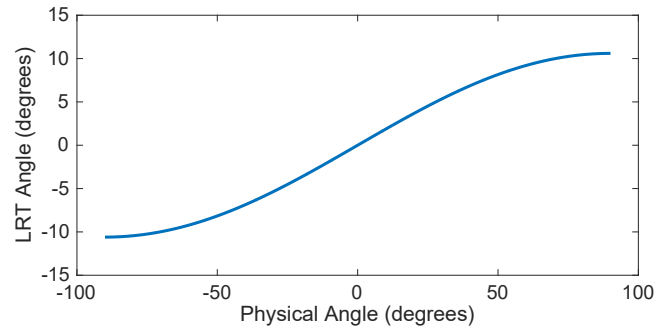
$$f_j(t) = \begin{cases} t/(T_{max} + T_{er}/2) & \text{for } T_{1,j} \leq t \leq T_{2,j} \\ 1 & \text{otherwise} \end{cases}, \quad (2)$$

and  $T_{max} = \max_{j=1..L}(T_{j,ref})$ ,  $T_{1,j} = \max(0, T_{j,ref} - T_{er}/2)$ ,  $T_{2,j} = \min(TT, T_{j,ref} + T_{er}/2)$  with  $TT$  the truncated RIR length from [13]. We use  $T_{er} = 3a/c$  where  $a$  is the array aperture and  $c$  is the speed of sound; this  $T_{er}$  is large enough to contain the direct sound and usually also the first reflection. A side effect of this procedure is that the direct sound peaks get more similar amplitudes across all RIRs.

The second adaptation we introduce is an array-geometry-aware correction of the detected main lobe peak of the direct sound's LRT response. The procedure maps the physical

<sup>4</sup>This is intended for the method to be compatible with real measurements.

<sup>5</sup>Strictly speaking, this is an attenuation for any distance  $r < 1$  m.



**Fig. 2:** Physical - LRT angle mapping from Eq.3, for  $LD = 0.1$  m,  $LD_{LRT} = 1/750$  m, and  $FS_{LRT} = 48$  kHz.

angle of arrival of the direct sound wavefront to an LRT-domain angle, i.e., the angle of the incoming wavefronts on the resampled stack. For transducers on a ULA, the physical angle of arrival is always  $\pm\pi/2$  ( $\pm$  depending on the relative ordering of the transducers), which corresponds to the maximum physically-valid angle on the LRT. However, this mapping<sup>6</sup> is used in more general cases (Section 3.2):

$$\theta_{LRT} = \text{sign}(\theta_{phys}) \text{atan2} \left( \left( LD.FS_{LRT} / c \left( (\tan(\theta_{phys} - \pi/2))^2 + 1 \right)^{1/2} \right), LD/LD_{LRT} \right), \quad (3)$$

where  $\theta_{LRT}$  is the LRT angle and  $\theta_{phys}$  is the physical angle of arrival of the wavefront to the microphone (both wrt the array center),  $LD$  is the physical transducer spacing on the array,  $LD_{LRT}$  and  $FS_{LRT}$  the transducer spacing and sampling frequency after resampling the RIR stack [13] Fig. 2 and  $\text{atan2}$  is the two-argument arc-tangent function. The result is then quantized to the angular grid  $\mathcal{A}$  of the LRT [13] and taken as the angular bin of the main lobe. The temporal bin of the main lobe is simply given by  $\sum_{j=1..L}(T_{j,ref})/L$ , and is also quantized to the sampled temporal grid. The value of the main lobe peak is then taken as the maximum LRT response inside the surrounding  $7 \times 7$  region<sup>7</sup>. The determined LRT peak is enforced at the early stages of the processing chain; it is substituted for the LRT peaks with the 5% highest amplitudes.

The third, trivial but important adaptation is to assume the microphone position is known via the known array geometry, thereby alleviating the need for SSL.

The fourth and last adaptation addresses the neighborhood suppression size used in the LRT processing [13]. As mentioned in Section 2.3, the early wavefronts suffer from near-field effects in our setup; this translates into considerably more spurious LRT peak response detections in this region. Therefore, for short ( $< 20$  cm) minimum microphone-loudspeaker

<sup>6</sup>This mapping shares similarities with the translation formula in [13, Section 4.4], albeit going from continuous (infinite) sampling to  $FS_{LRT}$  instead of going from  $FS_{LRT}$  to  $FS$ . A more advanced version, still retaining its general shape, would use a rounding function to account for the quantized grid on the stack, however this is not considered here as the LRT used in [13] allows for further interpolation between RIR stack pixels.

<sup>7</sup>This region and similar parameters are chosen empirically at our resampled spatial and temporal frequencies of 750 transducers/m and 48000 kHz [13].

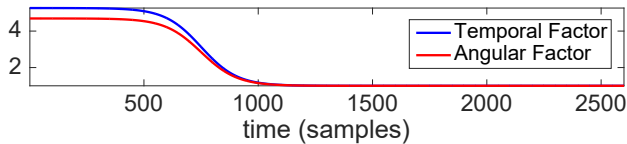


Fig. 3: Neighborhood suppression region size multiplication factors.

distances, we multiply the temporal and angular neighborhood suppression region sizes ( $Nl_x$  and  $Nl_y$  in [13], respectively) by two time-reversed sigmoid functions<sup>8</sup>:

$$1 + \frac{2T_{er}/3 - 1}{1 + \exp((T_{er} - n)/8T_{er})} \left( 1 - \frac{1}{1 + \exp((T_{er} - n)/8T_{er})} \right),$$

$$1 + \frac{2|\mathcal{A}| - 1}{1 + \exp((T_{er} - n)/8T_{er})} \left( 1 - \frac{1}{1 + \exp((T_{er} - n)/8T_{er})} \right),$$

for the temporal and angular dimensions, respectively, where  $n$  denotes the time in samples Fig. 3. The idea here is to gradually increase the suppression neighborhood going into the near-field region  $T_{er}$ . Moreover, we increase the image filter temporal width to 25, up from 15 samples (at 48 kHz) in [13], and we adjust the neighborhood suppression threshold  $T_r$  [13] to  $T_r + 55\%$  for the direct-sound peak specifically<sup>9</sup>. Finally, in contrast to our approach in [1, 13], we also allow reflection stack-lines to intersect the direct-sound stack-lines.

### 3.2. Setup-inspired and general improvements

In addition to the near-field adaptations, several improvements are introduced to the RGI method. Specifically, the first two improvements are made possible by knowledge of the  $\theta_{phys} - \theta_{LRT}$  angle mapping (Eq. 3).

The first improvement is the restriction, early in the processing chain, of the LRT peak response detection to physically-valid angles, i.e., angles that correspond to physical angles within  $[-\pi/2, \pi/2]$ . This is needed because, whereas it is not possible for sound waves to impinge on the array with bigger absolute angles, it is still theoretically possible for more-slanted but physically-invalid lines to appear on the RIR stack, and for the LRT to give a strong response to them. After all, the LRT is but a line detector in computer vision, with no such physical constraints.

The second and more significant improvement is a novel LRT region-spot-searching mode; which is promoted by the intra-array setup and which helps to achieve a more usable RL output. More specifically, the LRT peak-response detection is divided into three angular regions: 1)  $\theta_{phys} \leq -40^\circ$ , 2)  $-40^\circ \leq \theta_{phys} \leq 40^\circ$  and 3)  $\theta_{phys} \geq 40^\circ$  (all translated into LRT angles); for the lateral regions 1) and 3) we keep at most one salient LRT peak (if any), whereas for region 2) we keep the enforced (highest) direct sound peak from Section 3.1 in addition to at most the second- and third-highest peaks (if any, with a minimal time distance of  $a/c$  between these two latter). This step ensures that at most,

<sup>8</sup>These functions and their parameters are chosen empirically.

<sup>9</sup>This prevents erroneously discarding the LRT peak corresponding to the image microphone of the wall near the ULA.

and often exactly, four LRT peaks are detected (in addition to the direct peak); they correspond to the four walls of a rectangular room in 2D. This step solves the reflector selection problem left open (supervised) in [1, Section II-B]; it is effectively an automated reflector sifting mechanism which discards virtual (non-physical) reflectors, corresponding to second- and higher- order image microphones, and any other undesired reflector detections, e.g. the ceiling and floor detections when working in 2D with real measurements<sup>10</sup>. The boundaries between the regions make sense in the case of a ULA placed centrally near a wall in a shoebox room, as the image transducers corresponding to the side walls lie around  $\theta_{phys} \approx \pm\pi/2$ , and the image transducers corresponding to the front and back walls lie around  $\theta_{phys} \approx 0$ ; the angular ranges of the regions are intentionally chosen broadly in order to afford an error margin for LRT peak detection and to ensure robustness to different geometrical conditions, e.g., setups where the array is placed rotated wrt – instead of parallel to – the nearby wall.

The third improvement relates to an artifact of the LRT computation when slanting the RIR stack. The LRT can theoretically detect stack-lines with *negative* central time bin when they feature an angle  $|\theta_{phys}| > 0$ , such as a stack-line that intersects the array-center RIR in the stack at  $t = 0$  and that is rotated around this pixel. Accordingly, the LRT response is zero for  $\theta_{LRT} = 0, t < 0$ , but it follows a step<sup>11</sup> function pattern for  $|\theta_{LRT}| > 0, t < 0$ , especially so in the presence of noise or pre-ringing effects before the arrival of the direct sound. This step-response pattern can feign a genuine LRT response peak, especially in near-field scenarios, whereas it merely corresponds to the start of the data. Therefore, any LRT peaks within 7.5 LRT degrees and 15 samples of the artifact at ( $t = 0, \theta_{LRT} = 0$ ) are discarded, and any peaks with negative time bins are also discarded.

## 4. Performance evaluation

We perform two performance evaluations in this paper, one for TOA detection and labeling (Section 4.3) and one for RL (Section 4.4). The first evaluation gives information about how many of the reflectors are detected, whether correctly or incorrectly and how accurately (in terms of TOAs), as well as which reflectors are not detected. The second evaluation gives information about the RL error for the correctly detected physical reflectors.

### 4.1. Simulated setups and data sets

We re-used the same setups and performance evaluation frameworks as in [5, 13]; these consist of 7 different setups with different ULA configurations and room sizes; the only changes wrt our previous papers are the exclusion of real data and the move of the microphone positions from the cross pattern in the middle of the room (similar to [1, Fig. 11a] but in 2D) to the ULA itself, in line with the intra-array setup. To

<sup>10</sup>Both of these tasks are especially challenging in setups involving arrays with limited geometrical diversity.

<sup>11</sup>The start of this step corresponds to the start of the data at  $t = 0$  and occurs earlier for bigger absolute angles.

avoid exacerbating the already-challenging near-field effects, we only use the first three and the last three transducers on the ULA as microphones, i.e., we exclude the microphones around the array center. This means a total of  $7 \cdot 6 = 42$  independent RIR stacks for testing.

### 4.2. Methods under test

To elucidate the impact of each set of novelties on the basis algorithm from [1], we applied different versions of the method separately on the data sets, each version including a different set of adaptations/improvements:

- Version 0: basis algorithm from [1], used<sup>12</sup> in 2D.
- Version 1: Version 0 with the physical angle restriction and the improvements around  $t = 0$  (first and third improvements from Section 3.2).
- Version 2: Version 1 with all the near-field adaptations from Section 3.1.
- Version 3: Version 2 with the region-spot-searching mode (second improvement from Section 3.2).

We start by using our previous method unmodified from [1] (Version 0), which we then gradually but considerably expand: we first add changes independent of our intra-array setup (Version 1), then proceed to add intra-array-setup-specific adaptations (Version 2) to address the aforementioned challenges and then we finally add a major new feature to leverage the opportunities of the setup (Version 3). The distinctive advantage of Version 3 wrt Version 2 is the automatic, non-supervised discarding of second- and higher-order reflections.

### 4.3. TOA disambiguation metrics and results

The performance of the LRT-based TOA detection and labeling [13] was objectively assessed with three metrics: the true positive rate (TPR) indicating the percentage of detected TOA sets that match reference TOA sets, the number of false discoveries (FDs) of detected TOA sets that do not match reference TOA sets and the root mean square error (RMSE) between the correctly detected TOA sets' TOAs and their matched reference TOAs. Each detected TOA set was compared to all reference TOA sets, and counted as correct when a one-to-one match with an RMSE of 0.5 ms or less was found. The reference TOAs were retrieved from 2D simulations using the seventh-order image model [12]. Higher-order TOAs, and those beyond the truncation time  $TT$ , were not considered in the evaluation. All metrics were averaged across setups and microphone positions. Better performance is indicated by higher TPRs, fewer FDs and lower RMSEs. The same parameters as in [13] were used for the LRT processing.

The results (Table 1) show that the proposed adaptations result in similar robust performance as in [1]. Algorithm Versions 0 and 1 nearly fail given the intra-array setup, since they do not contain any of the adaptations addressing its challenges; this

<sup>12</sup>We use the same parameters as [1] with the exception of the new  $\hat{R} = 10$ .

**Tab. 1:** Obtained average TOA disambiguation performance metrics.

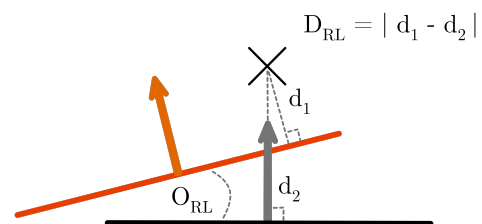
Order	All	1		2	
Alg. vers.	# of FDs	TPR %	RMSE $\mu$ s	TPR %	RMSE $\mu$ s
0	2.12	0	N/A	0	N/A
1	3.80	20.2	363.1	1.6	96.5
2	0.55	97.0	189.5	52.0	170.5
3	0.5	97.0	191.3	0.6	145.4

clearly motivates our proposed adaptations. More specifically, the LRT response involved in these versions is hardly usable given the aforementioned near-field effects, it only gives a splintered, diffuse response for the direct sound and a spurious artifact at  $t = 0$ ; these two diffuse responses often temporally, and less often angularly, coincide in our intra-array setup; they disproportionally overshadow any response from reflections, and the only way to avoid this is via the near-field adaptations. In both these algorithm versions, the method can at best (albeit with difficulty) detect the direct sound properly; this explains the low TPRs. The removal of the artifact at  $t = 0$  in Version 1 is inappropriate in these circumstances, as the spuriously detected artifact at  $t = 0$  would itself otherwise suppress many of the spurious peaks around the genuine-but-diffuse direct-sound response; this explains the jump in the number of FDs from Version 0 to Version 1.

Algorithm Versions 2 and 3 show remarkable and nearly-identical results (nearly-perfect first-order TPRs and a very low number of FDs). The main difference between these two versions is the nearly-complete discarding of second-order wavefront detections in Version 3; this actually fulfills the very purpose of this version: the automatic removal of second- and higher- order wavefront detections without compromising direct-sound and first-order wavefront detections (see Section 3.2).

### 4.4. RL metrics and results

To assess the accuracy of RL, the orientation error  $O_{RL} = \left| \arccos(\langle \mathbf{n}_r, \hat{\mathbf{n}}_r \rangle) \right|$  [15] between the true ( $\mathbf{n}_r$ ) and estimated ( $\hat{\mathbf{n}}_r$ ) reflectors' normal vectors was used; additionally, the offset  $D_{RL} = \left| \left| \langle \mathbf{n}_r, (\mathbf{m} - \mathbf{x}) \rangle \right| - \left| \langle \hat{\mathbf{n}}_r, (\mathbf{m} - \hat{\mathbf{x}}) \rangle \right| \right|$  [15] in terms of the distance of the true and estimated reflectors to the real microphone's true location  $\mathbf{m}$  was used, where  $\mathbf{x}$  and  $\hat{\mathbf{x}}$  represent points on the true and estimated reflectors, respec-



**Fig. 4:** Visual representation of RL error metrics. The black line, arrow and cross indicate a true reflector, its normal vector and image microphone position, while the red line and arrow indicate their estimated counterparts.

**Tab. 2:** RL performance metrics (average values followed by  $\pm$  the standard deviations), for algorithm versions 2 and 3.

Setup	Room size (m)	$D_{RL}$ (cm)	$O_{RL}$ ( $^\circ$ )
1	4.5x5	13.29 $\pm$ 14.13	7.45 $\pm$ 4.66
2	6x4	11.72 $\pm$ 9.20	7.66 $\pm$ 4.92
3	6x8.5	18.53 $\pm$ 26.58	7.37 $\pm$ 4.57
4	9x7.5	17.76 $\pm$ 24.17	7.32 $\pm$ 4.67
5	6x12	20.28 $\pm$ 39.26	7.35 $\pm$ 4.55
6	4.5x5	8.81 $\pm$ 14.37	6.09 $\pm$ 9.32
7	12.66x10.42	11.98 $\pm$ 8.66	3.00 $\pm$ 2.49

tively (Fig. 4). Only physical reflectors were considered, and the evaluation was done only for algorithm Versions 2 and 3; both versions gave the same metrics (Table 2), which were averaged across microphone positions (not averaged across setups). Lower metrics indicate better performance.

The results show degraded performance (+8.04 cm  $D_{RL}$  error and +4.89  $^\circ$   $O_{RL}$  error on average) wrt [1] (which shares identical but 3D-expanded configurations for setups 1-6); this is especially true for setups 4 and 5 (+11.38/12.09 cm  $D_{RL}$  errors and +5.09/5.04  $^\circ$   $O_{RL}$  errors, respectively). This shows that more adaptations are required to fully mitigate the near-field effects; however, it is worth noting that when the ULA is placed further away from the nearby wall and the room is larger (both conditions fulfilled in setup 7), angular error drastically decreases wrt other setups, and the distance error is also relatively lower; this is because the reflected wavefronts' near-field effects, which are not fully accounted for in the presented adaptations, are mitigated. The results are identical across algorithm versions 2 and 3 for the correctly-detected, reference-matched physical reflector detections involved; this is further evidence of the proper functioning of Version 3 (non-compromising of first-order reflections).

## 5. Conclusion

We presented an RGI method adapted for an intra-array transducer setup. The most important contribution in this respect is the adaptation to the near-field scenario. The second important contribution is a new mechanism for selectively sifting peak responses in the LRT domain by spot-searching in predetermined-but-broad regions; this automates the final reflector selection without compromising performance. The results show significant improvements wrt the existing RGI method from [1] with intra-array setups, with correct labeling of up to 97% of first-order echoes, albeit with degraded RL performance wrt [1] with non-intra-array setups.

## 6. References

[1] Y. El Baba, A. Walther, and E. A. P. Habets, "3D room geometry inference based on room impulse response stacks," *IEEE Trans. Audio, Speech, Lang. Process.*,

vol. 26, no. 5, pp. 857 – 872, May 2018.

- [2] F. Antonacci, J. Filos, M. R. P. Thomas, E. A. P. Habets, A. Sarti, P. Naylor, and S. Tubaro, "Inference of room geometry from acoustic impulse responses," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 10, pp. 2683–2695, Dec. 2012.
- [3] P. Annibale, J. Filos, P. Naylor, and R. Rabenstein, "Geometric inference of the room geometry under temperature variations," in *Proc. Intl. Symp. on Control, Communications and Signal Processing*, May 2012, pp. 1–4.
- [4] L. Remaggi, P. J. B. Jackson, P. Coleman, and W. Wang, "Acoustic reflector localization: Novel image source reversion and direct localization methods," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 2, pp. 296–309, Feb. 2017.
- [5] Y. El Baba, A. Walther, and E. A. P. Habets, "Reflector localization based on multiple reflection points," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Budapest, Hungary, Aug. 2016, pp. 1458–1462.
- [6] H. Naseri and V. Koivunen, "Cooperative simultaneous localization and mapping by exploiting multipath propagation," *IEEE Trans. Signal Process.*, vol. 65, no. 1, pp. 200–211, Jan. 2017.
- [7] R. Schmidt, "A new approach to geometry of range difference location," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 8, no. 6, pp. 821–835, Nov. 1972.
- [8] J. Scheuing and B. Yang, "Disambiguation of TDOA estimation for multiple sources in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1479–1489, Nov. 2008.
- [9] A. Moore, M. Brookes, and P. Naylor, "Room geometry estimation from a single channel acoustic impulse response," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Sep. 2013, pp. 1–5.
- [10] I. Dokmanic, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proceedings of the National Academy of Sciences*, vol. 110, no. 30, pp. 12 186–12 191, 2013.
- [11] F. Ribeiro, D. Florencio, D. Ba, and C. Zhang, "Geometrically constrained room modeling with compact microphone arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1449–1460, Jul. 2012.
- [12] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.
- [13] Y. El Baba, A. Walther, and E. A. P. Habets, "Time of arrival disambiguation using the linear Radon transform," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 106–110.

- [14] A. Beck, P. Stoica, and J. Li, "Exact and approximate solutions of source localization problems," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1770–1778, May 2008.
- [15] J. Filos, A. Canclini, F. Antonacci, A. Sarti, and P. Naylor, "Localization of planar acoustic reflectors from the combination of linear estimates," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Aug. 2012, pp. 1019–1023.







# Abstract Reviewed Paper at ICSA 2019

Presented \* by VDT.

## Detection of Constant Phase Shifts in Filters for Sound Field Synthesis

Frank Schultz, Nara Hahn, and Sascha Spors

*University of Rostock, Institute of Communications Engineering*

*Email: {frank.schultz, nara.hahn, sascha.spors@uni-rostock.de}*

### Abstract

Filters with constant phase shift in conjunction with 3/6 dB amplitude decay per octave frequently occur in sound field synthesis and sound reinforcement applications. These ideal filters, known as (half) differentiators, exhibit zero group delay and 45/90 degree phase shift. It is well known that certain group delay distortions in electro-acoustic systems are audible for trained listeners and critical audio stimuli, such as transient, impulse-like and square wave signals. It is of interest if linear distortion by a constant phase shift is audible as well. For that, we conducted a series of ABX listening tests, diotically presenting non-phase shifted references against their treatments with different phase shifts. The experiments revealed that for the critical square waves, this can be clearly detected, which generally depends on the amount of constant phase. Here, -90 degree (Hilbert transform) is comparably easier to detect than other phase shifts. For castanets, lowpass filtered pink-noise and percussion the detection rate tends to guessing for most listeners, although trained listeners were able to discriminate treatments in the first two cases based on changed pitch, attack and roughness cues. Our results motivate to apply constant phase shift filters to ensure that also the most critical signals are technically reproduced as best as possible. In the paper, we furthermore give analytical expressions for discrete-time infinite impulse response of an arbitrary constant phase shifter and for practical filter design.

### 1. Introduction

Sound reproduction systems for large audiences are often equipped with vertical loudspeaker arrays which shall deliver an equally pleasant acoustic experience to the audience in terms of loudness, timbre and spatial impression. The reproduced wavefront can be steered and shaped towards the target region by applying delays and weights to the individual loudspeaker signals [1, 2]. Due to the coherence of these, the sound field typically exhibits a low-pass filter characteristic, and thus needs to be equalized by high-pass filtering the audio input signal. As pointed out in [3, Ch. 3], the same array processing framework is used in wave field synthesis (WFS), which has only seemingly a different goal, namely the physical reconstruction of a desired sound field. The loudspeaker signals for WFS are derived from a high-frequency approximation of the Kirchhoff-Helmholtz integral

equation [4, 5]. This enables a computationally efficient implementation of WFS which comprises of, similar to large-scale sound reproduction, delays and weights for the individual loudspeakers and an overall equalization filter.

According to the theory of WFS [4], the specification of the equalization filter depends on the geometry and shape of the loudspeaker array. The transfer function is  $i\omega$  for 3D scenarios where 2D arrays (e.g. spherical or planar) are used. In terms of signals and systems theory this constitutes an differentiator, exhibiting a slope of +6 dB per octave and a constant phase of 90°. For 2D scenarios using 1D arrays (e.g. circular and linear), the filter is given as  $\sqrt{i\omega}$ , constituting a half-differentiator [6, 7], where both the slope and phase are halved to +3 dB per octave and 45°, respectively.

In practical systems, where a continuous and infinite array cannot be used, the specification of the equalization filter has to be adjusted accordingly. The usage of a practical array

built from individual loudspeakers causes spectral fluctuations above the so-called spatial aliasing frequency [4]. Moreover, due to the finite extent of the array, the synthesized sound field exhibits a low frequency roll off. The high-pass filter characteristic of an ideal equalization filter thus should be flattened out at the highest and lowest frequencies in the spectrum, resulting in a high-pass shelving filter. The upper limit coincides with the spatial aliasing frequency and the lower limit is determined by the spatial extent of the array [8].

The digital equalization filter is typically realized either in a finite impulse response (FIR) or infinite impulse response (IIR) form. FIR type equalization filters are often designed as linear phase, while omitting the above mentioned constant phase ( $90^\circ$  or  $45^\circ$ ) [9, 10]. This results in synthesized sound fields exhibiting a negative phase shift ( $-90^\circ$  or  $-45^\circ$ ) compared to the desired reference sound field (apart from the group delay of the FIR filter). There are also a number of IIR type equalization filters where the constant phase spectrum is explicitly taken into account [11] or comes as a byproduct of the minimum phase characteristics of the desired magnitude spectrum [3, 12]. The improved physical accuracy in the synthesized sound field is well demonstrated in [12, Fig. 9–11].

The audibility of constant phase shifts can be regarded as special issue of the audibility of phase distortion and group delay distortion, cf. [13–18], often evaluated with allpass filters. From these works it is known, that audibility is strongly dependent of the signal’s waveform and spectrum and the amount of the group delay in the critical bands. Generally, sensitivity for phase/group delay distortions decreases with increasing frequency. For low frequency content a different pitch and for high frequency content ringing and different lateralization is reported for group delay distortions. The polarity of highly transient signals plays a role for the audibility. It was often shown, that training on phase/group delay distorted audio content increases the sensitivity to detect them.

To the authors’ knowledge to date, the perceptual impact of the constant phase shift has not been studied yet. It is of great interest whether the existence or absence of such a phase shift is audible, and in the special context of sound field synthesis, if this affects the authenticity of the synthesized sound fields. The paper discusses the signal processing fundamentals of discrete-time constant phase shift in Sec. II. In Sec. III a listening test is presented for selected audio content and phase shifts to initially evaluate the audibility of constant phase shifts. Sec. IV concludes the paper.

## 2. Constant Phase Shifter

A constant phase shifter, also known as fractional Hilbert transformer [19], refers to a filter that applies a (frequency independent) constant shift  $\varphi$  to the spectrum of an input signal. This section introduces the time and frequency representations of discrete-time constant phase shifters for aperiodic and periodic signals. Practical implementations for the respective cases are also discussed. Since the primary interest of the present study is the audibility of a phase shift, the main consideration is the accuracy of the constant phase shifter in terms of its magnitude and phase response. Computational

cost and algorithm optimization are less of a concern.

### 2.1. Aperiodic Signals

The transfer function of a constant phase shifter in the discrete time Fourier transform (DTFT) domain reads

$$H(e^{i\Omega}) = \begin{cases} e^{+i\varphi}, & 0 < \Omega < \pi \\ e^{-i\varphi}, & -\pi < \Omega < 0 \\ \cos \varphi, & \Omega = 0, \pi \end{cases} \quad (1)$$

where  $\Omega = \frac{2\pi f}{f_s}$  denotes the normalized angular frequency for the sampling rate  $f_s$ . By exploiting Euler’s formula, (1) can be decomposed into

$$H(e^{i\Omega}) = \cos \varphi - \sin \varphi \cdot H_H(e^{i\Omega}), \quad (2)$$

with  $H_H(e^{i\Omega}) := -i \cdot \text{sgn}_\Omega$  denoting the transfer function of the Hilbert transformer [20]. Notice that the Hilbert transformer can be regarded as a constant phase shifter of  $\varphi = -\frac{\pi}{2}$ . Since  $H_H(e^{i\Omega})$  is free of DC bias [20, Sec. 4.2], the magnitude response of a constant phase shifter is unity at all frequencies but  $\Omega = 0, \pi$ , as given in (1).

The discrete-time impulse response of a constant phase shifter is obtained by computing the inverse DTFT of  $H(e^{i\Omega})$ ,

$$h[n] = \begin{cases} \cos \varphi, & n = 0 \\ 0, & n \neq 0 \text{ and even} \\ -\frac{2}{n\pi} \sin \varphi, & n \text{ odd.} \end{cases} \quad (3)$$



Analogous to (2), it comprises of two components,

$$h[n] = \cos \varphi \cdot \delta[n] - \sin \varphi \cdot h_H[n] \quad (4)$$

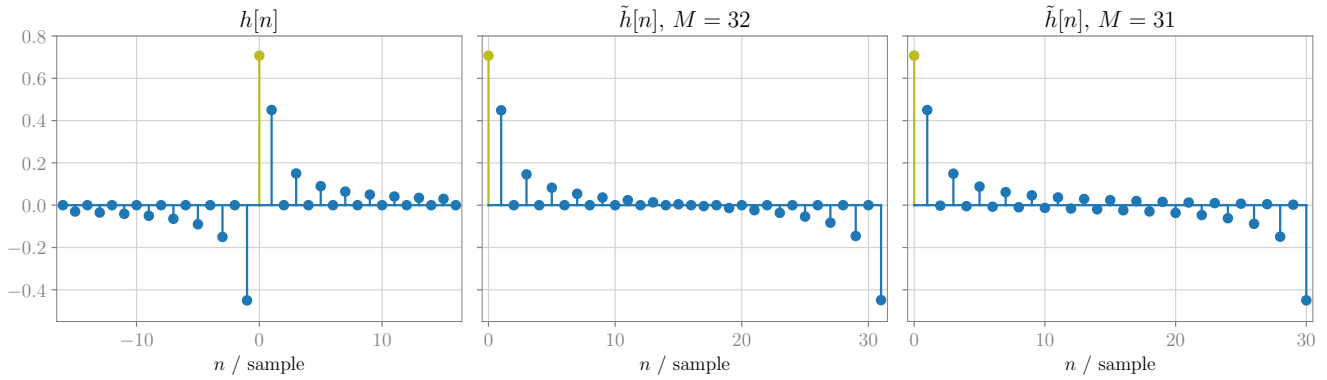
where

$$h_H[n] = \begin{cases} 0, & n \text{ even} \\ \frac{2}{n\pi}, & n \text{ odd.} \end{cases} \quad (5)$$

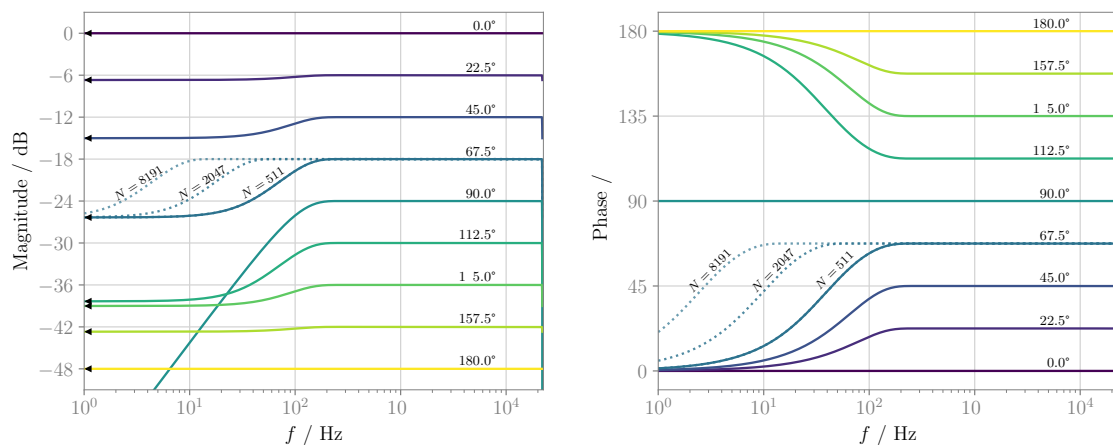
denotes the discrete-time impulse response of the Hilbert transform filter [21, Eq. (11.65)]. The constant phase shift of an input signal is thus a linear combination of the input itself and its Hilbert transform, weighted with  $\cos \varphi$  and  $-\sin \varphi$ , respectively.

In Fig. 1(left), the impulse response  $h[n]$  for  $\varphi = -\frac{\pi}{4}$  is depicted. It can be seen that the impulse response is of infinite length and not causal. The coefficients for  $n \neq 0$ , indicated by , exhibit odd symmetry with respect to  $n = 0$  and add up to 0. The DC and  $\frac{f_s}{2}$  gain are thus solely determined by  $h[0] = \cos \varphi$ , indicated by  and consistently given in (1).

Filtering with the impulse response  $h[n]$  is not feasible in practice due to the infinite extent of  $h[n]$ . An FIR constant phase shifter can be built by applying a finite window to  $h[n]$ , known as windowing method [21, Sec. 7.2]. Considering the decay of the coefficients for  $|n| \rightarrow \infty$ , it is a natural choice to truncate the impulse response symmetrically with respect to  $n = 0$ , which leads to an even-order (odd-length) FIR filter. Since  $h[n]$  vanishes for even  $n \neq 0$ , the FIR length  $N$  has to satisfy  $N \bmod 4 = 3$  (i.e.  $N = 3, 7, 11, 13, \dots$ ). Using a tapering window is advantageous as it smooths out the



**Fig. 1:** Left: Impulse response of a constant phase shifter ( $\varphi = -45^\circ$ ) described in (3). Center & Right: Impulse responses of periodic constant phase shifter ( $\varphi = -45^\circ$ ) for even and odd period  $M$ , described in (8) and (9), respectively. The coefficients  $h[0] = \tilde{h}[0] = \cos \varphi$  are indicated by  $\blacktriangledown$ . The coefficients for  $n \neq 0$  indicated by  $\blacktriangledown$  show the Hilbert transform part of the impulse response.



**Fig. 2:** Frequency responses of constant phase shifters (left: magnitude, right: phase,  $f_s = 44.1$  kHz). The FIR filter coefficients are obtained by using the windowing method (filter length of  $N = 511$  for all angles and  $N = 511, 2047, 8191$  for  $\varphi = 67.5^\circ$ ). The magnitude responses are depicted with 6 dB offsets. The triangles  $\blacktriangleleft$  indicate the DC gain  $\cos \varphi$ . The phase responses are obtained by compensating the group delay  $\tau = \frac{N-1}{2} \frac{1}{f_s}$ , i.e. multiplying the complex exponential  $e^{i2\pi f \tau}$  to the spectra.

ripples in the frequency domain. Due to the non-causality of the filter, a constant group delay of  $\tau = \frac{N-1}{2} \frac{1}{f_s}$  is introduced additionally to the desired property given by (1).

Exemplary frequency responses of FIR constant phase shifters ( $\varphi = 0^\circ, \dots, 180^\circ$ ) are shown in Fig. 2. The FIR coefficients are obtained by applying a Blackman window to  $h[n]$  in (3). It can be seen that the desired magnitude and phase responses are achieved only within a limited frequency band. For  $\varphi \neq 0^\circ, 180^\circ$ , the magnitude responses typically attenuate at low and high ends ( $f = 0$  and  $f = \frac{f_s}{2}$ , respectively) and converge to  $\cos \varphi$ . Due to the logarithmic frequency axis, the behavior at  $f = \frac{f_s}{2}$  is not clearly visible, though. For  $\varphi \neq 90^\circ$ , the phase responses are inaccurate at both ends and gradually tend to  $0^\circ$  or  $180^\circ$ . The constant phase shifter for  $\varphi = 90^\circ$  exhibits an ideal phase response but the most distorted magnitude response among other phase angles. Accurate frequency responses are observed for  $\varphi = 0^\circ, 180^\circ$  where the filters are integer delays ( $\tau = \frac{N-1}{2} \frac{1}{f_s}$ ) with non-inverted and inverted polarity. As indicated by dotted lines ( $\varphi = 67.5^\circ$ ) in Fig. 2, the spectral distortions can be suppressed by increasing the FIR filter order, which comes at the expense of an increased group delay.

## 2.2. Periodic Signals<sup>1</sup>

Consider an  $M$ -periodic signal  $\tilde{s}[n] = \tilde{s}[n + M]$  and the generating signal  $s[n]$  which coincides with  $\tilde{s}[n]$  for the period  $n = 0, \dots, M-1$  and vanishes elsewhere. Then the periodic signal can be represented as a shifted sum of  $s[n]$ ,

$$\tilde{s}[n] = \sum_{\mu=-\infty}^{\infty} s[n] *_{n} \delta[n - \mu M], \quad (6)$$

where  $*_{n}$  denotes the linear convolution with regard to  $n$ . The constant phase shift of  $\tilde{s}[n]$  reads

$$\tilde{y}[n] = \tilde{s}[n] *_{n} h[n] = s[n] *_{n} \underbrace{\sum_{\mu=-\infty}^{\infty} h[n - \mu M]}_{\tilde{h}[n]}, \quad (7)$$

meaning that the generating function  $s[n]$  is convolved with an infinite sum of shifted impulse responses denoted by  $\tilde{h}[n]$ . A closed form expression for  $\tilde{h}[n]$  can be obtained by

<sup>1</sup>The derivation in this subsection is adopted from [20, Sec. 1.9 and Sec. 4.6].

substituting (4) for  $h[n]$  and exploiting the series expansion of the cotangent function [22, Eq. (4.3.91)], reading

$$\tilde{h}[n] = \begin{cases} \cos \varphi, & n = 0 \\ 0, & n \neq 0 \text{ and even} \\ -\frac{2 \sin \varphi}{M} \cot\left(\frac{\pi n}{M}\right), & n \text{ odd,} \end{cases} \quad (8)$$

for even  $M$ , and

$$\tilde{h}[n] = \begin{cases} \cos \varphi, & n = 0 \\ -\frac{\sin \varphi}{M} \cot\left(\frac{\pi(n+M)}{2M}\right), & n \neq 0 \text{ and even} \\ -\frac{\sin \varphi}{M} \cot\left(\frac{\pi n}{2M}\right), & n \text{ odd,} \end{cases} \quad (9)$$

for odd  $M$ . Exemplary impulse responses are shown in Fig. 1(center, right) for  $\varphi = -45^\circ$ .

A periodic repetition of a signal in the time domain is equivalent to a sampling of the spectrum in the DTFT domain [21, Sec. 8.4]. The discretized DTFT spectrum then constitutes the discrete Fourier transform (DFT) spectrum [21, Sec. 8.5], for which periodicity of the signal and the spectrum are inherent. The DFT coefficients for even  $M$  read

$$H[k] = H(e^{i\frac{2\pi}{M}k}) \quad (10)$$

$$= \begin{cases} e^{+i\varphi}, & k = 1, \dots, \frac{M}{2} - 1 \\ e^{-i\varphi}, & k = \frac{M}{2} + 1, \dots, M - 1 \\ \cos \varphi, & k = 0, \frac{M}{2} \end{cases}$$

where  $H[k]$  denotes the DFT of  $\tilde{h}[n]$ . Note that (8), (9), and (10) are analytic representations of the constant phase shifter with no approximations involved. An ideal constant phase shift is therefore feasible as far as periodic discrete-time signals are concerned.

In practice, a constant phase shift of a periodic signal can be computed efficiently in the DFT domain,

$$y[n] = \frac{1}{M} \sum_{k=0}^{M-1} S[k] H[k] e^{i\frac{2\pi}{M}kn}, \quad (11)$$

for  $n = 0, \dots, M - 1$ , where  $S[k]$  denotes the DFT of  $s[n]$ . This constitutes a circular convolution of  $s[n]$  and  $\tilde{h}[n]$ , and  $y[n]$  exhibits a temporal aliasing which constitutes the desired result, as shown in (7). Finally, the periodic signal  $\tilde{y}[n]$  is constructed with the generating signal  $y[n]$ , in the same way as (6).

### 3. Listening Experiment

We aim at investigating, if audio content treated with a constant phase shift filter can be perceptually discriminated from the original signals. This section discusses the design, procedure and analysis of the conducted listening experiment, related to this question.

#### 3.1. ABX Test Framework

The discrimination performance was tested with the highly sensitive two alternatives, forced choice ABX test. Stimuli A and X were both randomly assigned to either the reference (original) or the treatment (phase shift), subsequently ensuring

that B contains the other stimulus than A. According to the ABX test paradigm, test subjects were asked to assign either X=A or X=B after thorough, non-time-limited comparison of A, B and X.

No looped playback or instantaneous stimulus switching with crossfade or fast fade-out/fade-in could be utilized, since treatment detection would have become a trivial task based on the resulting artifacts (i.e. clicks for fade-out/in, phasing for crossfade). Instead, the stimuli—always (re)-started from the beginning—had to be manually started and stopped by the test subjects. This ensures artifact free playback, although with higher interaction. Moreover, the requirements led to some modifications of the utilized webMUSHRA test framework [23], which by default intends seamless switching and looping by fading out/in.

#### 3.2. Audio Content

The 4 monaural audio contents

- three square wave burst signals, each: 50 Hz, 200 ms on including 40 ms  $\sin^2$ -fade in and out, 300 ms off. Fourier series synthesis of the harmonics 1, 3, ..., 19, modal windowing (Kaiser,  $\beta = 4$ ) of the Fourier coefficients. total length 1.5 s @ 120 bpm, periodicity assumed / DFT filtering
- pink noise<sup>2</sup>, lowpass filtered (4th order Butterworth, cut frequency 300 Hz), length 2.35 s, non-periodicity assumed / FIR filtering
- castanets<sup>3</sup>, length 2 s, periodicity assumed / DFT filtering
- Hotel California, Eagles, Hell freezes over, 1994, Geffen, stereo version mixdown to mono, time range 0:44.938 - 0:48.142, non-periodicity assumed / FIR filtering

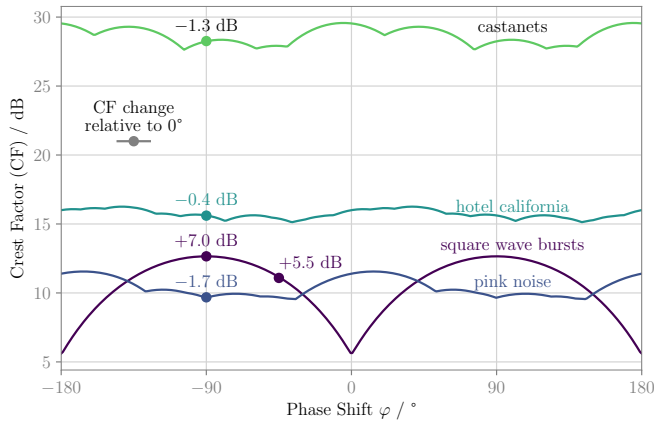
were chosen to create the 5 treatments

- I square wave burst with  $\varphi = -90^\circ$
- II square wave burst with  $\varphi = -45^\circ$
- III lowpass filtered pink noise with  $\varphi = -90^\circ$
- IV castanets with  $\varphi = -90^\circ$
- V Hotel California with  $\varphi = -90^\circ$ ,

according to the following considerations: It is known that human hearing is sensitive to group delay variations of low frequency square wave bursts [16], which is assumed to hold for constant phase shifts as well. To evaluate a potential detection of phase shifts for highly transient audio material and to check for potential detection of pre-/postringing due to filtering of such, castanets were included. Phase alignment in noise signals is highly random. It was assumed that human hearing is sensitive to a varied phase structure of noise with a low frequency spectrum. Furthermore, a musical record with a percussion sequence—commonly considered as very high fidelity production (mixing: E. Scheiner, mastering: T.

<sup>2</sup>generated by Voss-McCartney algorithm:  
<https://github.com/AllenDowney/ThinkDSP/blob/master/code/voss.ipynb>

<sup>3</sup>anechoic version of EBU SQAM CD track 27:  
<https://iaem.at/Members/frank/sounds/castanets-dry>



**Fig. 3:** Crest factor over phase shift for the 4 audio contents used in the listening test.

Jensen)—was included. For that content, it was assumed that a constant phase shift might be audible in the decay of the drumheads or/and in the transients.

For the treatments I, II and IV signal periodicity is assumed and thus the ideal phase shift filter via DFT (Sec. 2.2) was applied. To create treatments V and III, no signal periodicity was assumed for the whole musical piece as well as for the generated pink noise raw material of 6 minutes duration. Thus FIR filtering according to Sec. 2.1 was realized. Considering the audio contents as rectangularly windowed signals of infinite duration, the filter order of 3963530 ( $\approx 90$  s!) ensures that linear convolution of the chosen excerpt of Hotel California is complete. The resulting magnitude ripple of the Blackman windowed FIR is negligible for the relevant reproduction bandwidth. Since the pink noise length can be arbitrarily set, the same FIR filter was utilized for consistence.

### 3.2.1. Crest Factor Discussion

An ideal constant phase shifter does not alter the spectrum of the input signal, nor does it introduce any group delay. One noticeable technical change concerns the waveform, quantifiable in terms of the crest factor (CF),

$$CF = 20 \log_{10} \left( \frac{|s|_{\text{PEAK}}}{s_{\text{RMS}}} \right) \text{ in dB.} \quad (12)$$

During the preparation of this study, it was speculated that an increase or decrease of the CF might lead to detectable cues (if there are any) for a phase shift.

Figure 3 depicts the CF of the chosen audio contents for varying phase shifts  $\varphi \in [-180^\circ, 180^\circ]$ . For aperiodic signals (pink noise and Hotel California), the phase shift is applied to the entire piece but the CF is evaluated only for the selected part which is used in the listening experiment. Notice that the CF is  $180^\circ$ -periodic. This is because a phase shift of  $180^\circ$  reverses the polarity of the signal and the CF remains unchanged. Among other stimuli, the square wave burst shows the highest variation of approximately 7 dB. Note, however, that due to the silence between the bursts and the temporal shaping by the fade-in/-out window, the CF deviates from that of a continuous square wave (which is theoretically

$CF = 0$  dB for  $\varphi = 0^\circ, \pm 180^\circ$  and  $CF = \infty$  for  $\varphi = \pm 90^\circ$ ). The CF of castanets is about 15 to 25 dB higher than other stimuli because of its fast attack/decay, very short sustain, and relatively long silence.

As mentioned in Sec. 1, phase angles of  $-90^\circ$  and  $-45^\circ$  are particularly of our interest due to the relation with the equalization filter in WFS. Except for square waves, a phase shift of  $-45^\circ$  is barely detectable for most of the audio materials that was tested in informal listening. Therefore,  $-90^\circ$  is predominantly tested in this study whereas  $-45^\circ$  is included only for square wave bursts. In Fig. 3, the CF change of the phase shifted stimuli relative to the original signal ( $\varphi = 0^\circ$ ) is annotated above the filled circles  $\bullet$ .

### 3.2.2. Audio Signal Processing and Rendering

Each reference audio content (except castanets) was loudness calibrated to -23 LUFS [24]. The according calibration gain was also applied to the associated phase shifted stimuli. Since the loudness measure of [24] is suboptimal for castanets, perceptually motivated re-calibration to -35 LUFS was pursued in order to better match playback level with the other audio contents. Non-dithered 24 Bit, 44.1 kHz PCM wav-files were rendered for all required stimuli, carefully monitoring that amplitude clipping—as undesired artifact blended with the phase shift—does not occur.

### 3.3. ABX Test Statistics Considerations

All test subjects were to rate the 5 different treatments created from the 4 audio contents in mixed, randomized order and—except for the two lead authors—without preliminary training or other preconditioning with respect to the research question.

For each of the 5 treatments 25 ABX trials had to be rated, resulting in  $5 \cdot 25 = 125$  judgments per test subject aiming at evaluation of individual detection rates in the first instance. Thus, with underlying Binomial distribution model [25, 26], these quantities originate from intended one-side tail hypothesis testing of the  $\mathcal{H}_0(p_{\text{detect}} = 0.5)$  using Bonferroni correction to a target rejection level  $\alpha = 0.05$ , a target test power  $1 - \beta = 0.95$  and an effect size of  $g = 0.4$ , which was determined from preliminary test results using square wave bursts, achieving detection probabilities of about  $p_{\text{detect}} = 0.9$ .

Assuming independence of all collected ratings, contingency tables of detection frequencies can be statistically evaluated with underlying Chi-Square ( $\chi^2$ ) distribution model as post hoc tests.

### 3.4. Experiment Procedure

The listening test was conducted in our loudspeaker array lab with 0.3 s mean RT60 and about 40 dB(A)<sub>Leq</sub> sound pressure level (SPL) of background noise. A large monitor, a keyboard and a mouse were set up on a table, where test subjects took seat in the middle of the lab during the experiment. The browser based ABX GUI of the webMUSHRA software was hosted on an Apple Mac Mini connected to an RME Fireface UC.

Playback was presented diotic (i.e. same signal for both ears) using an electrodynamic, circumaural, open headphone Sennheiser HD 800. Playback level was settled such that for a mono pink noise signal with -23 LUFS loudness



( $-7.8 \text{ dB}_{\text{TruePeak}}$ ,  $-21.3 \text{ dB}_{\text{RMS}}$ , i.e.  $13.5 \text{ dB}_{\text{CF}}$ ) sound pressure levels (SPLs) of  $72.7 \text{ dB(A)}_{\text{Leq}} / 87.7 \text{ dB(C)}_{\text{Peak}}$  were measured for the left and the right channel using a G.R.A.S. headphone-to-ear coupler and a calibrated Brüel & Kjaer SPL meter according to the IEC 60318 standard.

12 test subjects (4 female, 8 male) took part in the listening experiment. Except one light tinnitus afflicted, all others reported normal hearing. Test subjects' age distribution is given as  $\mu_a = 29.8$ ,  $\sigma_a = 6.5$  years with the percentiles  $a_{0.05} = 22.6$ ,  $a_{0.25} = 25$ ,  $a_{0.5} = 28$ ,  $a_{0.75} = 32$ ,  $a_{0.95} = 40.5$  years. About half of the listening test panel consists of music production experts and (future) professional musicians (classical instrument students). The other half recruits from research related, untrained listeners, occasionally without prior experience in performing listening tests.

For familiarization of the specific ABX GUI, ratings on full band pink noise with 1 dB level difference were performed by each test subject prior to the actual listening test. This procedure was accompanied by written operational instructions and remarks, indicating that the differences to be detected in the actual listening test might be very subtle and will potentially differ in character compared to that of the training session. Participants (except the two authors) were left completely uninformed with respect to the signal manipulation method, thus expecting unbiased strategies for the detection of differences.

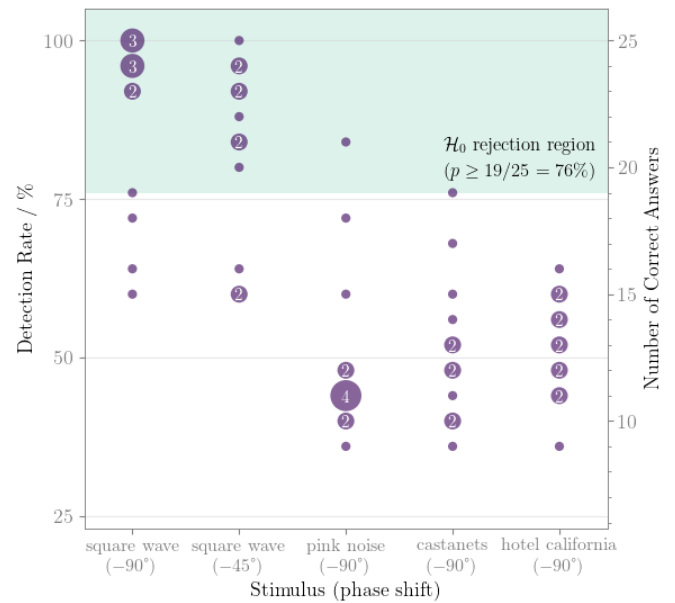
After many pre-listening experiments, we decided against playback of all 25 trials per each treatment in one sequence. This regularly resulted in an extremely tedious task, that should only be considered for few well trained and extraordinary performing test subjects. Since, in the first instance, we intended to find effect sizes  $g$  for a rather untrained, non-preconditioned test panel, we set up playback for a mixed randomized sequence of all 125 trials, divided in 4 parts ( $35 + 3 \cdot 30$ ) with longer intermissions. Thus, the results to be presented in the next section, should be considered for this conditioning.

### 3.5. Results

#### 3.5.1. Tests on Binomial Distribution

The detection rates (correct discrimination between original and phase shifted version) of all test subjects for all treatments are shown as a scatter plot in Fig. 4. For the  $-90^\circ$  phase shifted square wave burst, except of three subjects, all other perform with very high, statistically significant detection rates between 76% and 100%. For the  $-45^\circ$  phase shifted square wave burst, except of the same three subjects, all other show very high, statistically significant detection rates between 80% and 100%.

The  $-90^\circ$  phase shifted pink noise treatment exhibits only one result for which guessing can be excluded by statistical significance. This was produced by one of the preconditioned, trained authors. The majority of detection is slightly below guessing. About the same situation, however with larger spread of the rates, can be observed for the  $-90^\circ$  phase shifted castanets treatment. The one statistically significant result was achieved by another trained, preconditioned author. All other detection rates fail to reject that guessing took place.



**Fig. 4:** Detection rates in percent and number of correct detections of all test panelists. Small dots without numbering indicate a performance of a single test subject, whereas increasing dots with numbering indicate same performance for multiple test subjects. Results within the shaded area indicate that guessing is unlikely with statistical significance.

For the musical content, i.e. the percussion sequence in Hotel California, no statistical significance can be reported. Detection rates spread around the guessing rate ranging from 36% to 64%.

#### 3.5.2. Tests on Chi-Square Distribution

The  $\mathcal{H}_0$  (correct and incorrect discrimination occur with equal frequency) is tested with the  $\chi^2$  test per treatment considering all judgments. Bonferroni correction for total of 5 treatments to a target  $\alpha = 0.05$  was considered. The results indicate that this  $\mathcal{H}_0$  can be rejected with very high statistical significance for the two square wave bursts, but not for the other three treatments.

The  $\mathcal{H}_0$  (detection performance of two treatments is independent) is tested with the  $\chi^2$  test of the contingency table built from two treatments considering all judgments. Bonferroni correction for the 10 possible pairwise comparisons to a target  $\alpha = 0.05$  was considered. The results indicate that this  $\mathcal{H}_0$  can be rejected for the pairs I vs. III, IV, V and II vs. III, IV, V with very high statistical significance and odds ratios (ORs) between 4.5 to 6.5. For the pairs I vs. II and III vs. IV, V and IV vs. V we fail to reject this  $\mathcal{H}_0$ . For pairwise comparison I vs. II the OR  $\approx 1.4$  indicates a very small (however statistically not significant,  $p = 0.17$ ) tendency for a different performance. For the other pairs OR  $\approx 1$  indicates very comparable rating performance.

#### 3.5.3. Rating Durations

Test subjects took between 70 and 120 minutes in total for the whole listening experiment, including intermissions. One test subject asked to split the test onto two days for improved power of concentration.

The webMUSHRA software captures data for rating durations of all trials. This duration is defined as the time interval from initiating the GUI for the actual trial up to submitting the result and moving on with the next trial. Thus, this duration measure includes all (potentially longer) individual intermissions of a test subject. Pure playback duration of A, B, X per trial, which is considered a more useful measure, is unfortunately not available. We thus report the percentiles in Table 1 over all rating durations  $t$  in seconds without further statistic evaluation. However, the table easily reveals that the median and the interquartile range—that should not overly affected by longer intermission intervals—increase following the treatment sequence I to V.

treatment	$t_{0.05}/s$	$t_{0.25}/s$	$t_{0.5}/s$	$t_{0.75}/s$	$t_{0.95}/s$
I sq -90	8.1	12.4	17.7	29.2	60.9
II sq -45	8.9	15.3	23.1	39.8	95.2
III pn -90	13.7	19.8	28.7	47.4	98.6
IV cas -90	11.5	20.2	30.0	48.2	109.7
V hc -90	12.6	24.4	35.9	52.5	104.2

**Tab. 1:** Percentiles of the rating duration per ABX trial.

### 3.5.4. Qualitative Statements

Test subjects were asked for handwritten qualitative statements (e.g. detection cues, artifacts) during the listening experiment. Since this was handled as unforced add-on not all panelists reported back. However, the received statements are highly valuable and can be summarized as follows.

Treatments using square wave bursts were comparably very easy to detect, and comparing both square waves, the  $-90^\circ$  treatment was much easier to detect than the one with  $-45^\circ$ . Most often pitch shifts were used as cues, but also changes in envelopment and dispersion were reported.

The detection of pink noise treatments was reported as very demanding. Here test subjects indicated changed pitch, roughness, sharpness, subbass structure, melody and ambience as cues. For castanets test subjects reported a hard time for detection and admitted pure guessing very often using either the pitch or/and the characteristics of the very first transient (change of attack, punch and crispness) to discriminate treatment and reference. No pre-/postringing artifacts were stated for the castanets.

For the percussion sequence (Hotel California) most reports agreed on pure guessing. However, test subjects also reported there to use changed decay of drumheads and changed pitch as cues as well as smeared transients.

One test subject reported that A,B and X were perceived with different pitches for the square wave bursts. Here, due to the forced choice design, the achieved detection rates must fail to reject  $\mathcal{H}_0$ , which was confirmed post-hoc.

## 4. Conclusion

This study exhibits explorative character to firstly evaluate the audibility of constant phase shifts. For signals with rather complex structure, arbitrary constant phase shifts can only be realized with digital signal processing. Based on the Hilbert transform (i.e. the special case of  $-90^\circ$  constant phase

shift with unit magnitude), for which the infinite impulse responses are well known in continuous-time and discrete-time signal domain, this paper introduces the discrete-time infinite impulse response for an arbitrary constant phase shifter. For practical implementations an FIR design is proposed with special attention to retain unit magnitude. Furthermore, for periodic signals a periodic convolution is discussed. Under the periodicity assumption, hereby the ideal phase shift filter without any approximations or limitations can be applied. The periodic convolution can be computed within DFT domain with high performance, for which the spectrum of the constant phase shifter is given.

The results of the conducted listening experiment can be referred to the following deductions. Untrained, unconditioned listeners that have been repeatedly confronted with multiple low frequency square wave bursts, lowpass filtered noise, a transient castanet rhythm and a percussion sequence in a randomized sequence, in general show different detection performances of applied constant phase shifts.

The majority of listeners was able to discriminate constant phase shifts of  $-90^\circ$  and  $-45^\circ$  for square wave bursts with comparably little demand and very high detection rate. A 100% detection rate was achieved by musical experts. These findings are according to the known results with respect to other low frequency group delay distortion of square waves.

For pink noise the majority of listeners is not able to detect constant phase shift treatments of  $-90^\circ$ . However, the results indicate that by musical background and audio expertise higher detection rates can be achieved, that might be tested for statistical significance by a more sensitive test design. An adapted effect size of about  $g = 0.2$  to  $0.25$  seems to be reasonable for this. The same observation and conclusion holds for the  $-90^\circ$  constant phase shift of the castanets signal. For both signals trained listeners are able to detect the treatment with statistical significance. The initial guessed effect size can be well assumed for both stimuli.

All listeners were not able to detect constant phase shift of  $-90^\circ$  for the sequence containing percussion material with full audio bandwidth. Here, highest judgment demand was reported by qualitative statements, very often admitting pure guessing. The comparably longest rating durations might reflect this fact as well. This insensitivity might be due to complex spectrum and full audio bandwidth, compared to the other used audio contents. An adapted effect size of about  $g = 0.15$  seems to be an appropriate choice for a more sensitive test design (e.g. addressing significant detection rates  $\frac{\geq 69 \text{ correct}}{119 \text{ total}}$  of single treatment judgment for  $\alpha = \beta = 0.05$ ).

Although, here only shown in an ABX comparison scenario and not yet for musical contents, we cannot fully exclude that very critical, trained listeners are able to detect constant phase shifts of well known references even in a non-comparison task as well. Considering this circumstance and the current listening experiment results, it appears advisable to apply constant phase shift filters in 2.5D and 3D sound field synthesis applications to perfectly guarantee that potentially audible phase shift artifacts will not occur.

## Open Science

This project is following the open science paradigm. Please find all relevant code and data in the related git repository<sup>4</sup>. The DOI <https://doi.org/10.5281/zenodo.3383286> is directly related to the repository's state when submitting the paper. The repository includes Jupyter notebooks for signal processing calculus and statistical evaluation, stand-alone Python code to create all presented figures, the tex source code for the paper and the related talk, the raw data from the listening experiment, the code modifications of the used ABX software as well as the ABX configuration files. Besides the copyrighted piece of music, all other audio content used for the listening experiment is freely available.

## 5. References

- [1] D. G. Meyer, "Digital control of loudspeaker array directivity," *J. Audio Eng. Soc.*, vol. 32, no. 10, pp. 747–754, 1984.
- [2] —, "Multiple-beam, electronically steered line-source arrays for sound-reinforcement applications," *J. Audio Eng. Soc.*, vol. 38, no. 4, pp. 237–249, 1990.
- [3] F. Schultz, "Sound field synthesis for line source array applications in large-scale sound reinforcement," Ph.D. dissertation, University of Rostock, 2016.
- [4] S. Spors, R. Rabenstein, and J. Ahrens, "The theory of wave field synthesis revisited," in *Proc. 124th Audio Eng. Soc. Conv.*, Amsterdam, 2008.
- [5] F. Zotter and S. Spors, "Is sound field control determined at all frequencies? How is it related to numerical acoustics?" in *Proc. 52nd Int. Conf. Audio Eng. Soc. (AES)*, 2013.
- [6] C.-C. Tseng, S.-C. Pei, and S.-C. Hsia, "Computation of fractional derivatives using Fourier transform and digital FIR differentiator," *Signal Processing*, vol. 80, no. 1, pp. 151–159, January 2000.
- [7] B. Krishna, "Studies on fractional order differentiators and integrators: A survey," *Signal Processing*, vol. 91, no. 3, pp. 386–426, March 2011.
- [8] S. Spors and J. Ahrens, "Analysis and improvement of pre-equalization in 2.5-dimensional wave field synthesis," in *Proc. 128th Audio Eng. Soc. Conv.*, London, 2010.
- [9] H. Wierstorf, "Perceptual assessment of sound field synthesis," Ph.D. dissertation, Technische Universität Berlin, 2014.
- [10] F. Winter, "Local sound field synthesis," Ph.D. dissertation, University of Rostock, 2019 (to be appear).
- [11] C. Salvador, "Wave field synthesis using fractional order systems and fractional delays," in *Proc. 128th Audio Eng. Soc. Conv.*, London, 2010.
- [12] F. Schultz, V. Erbes, S. Spors, and S. Weinzierl, "Derivation of IIR-pre-filters for soundfield synthesis using linear secondary source distributions," in *Proc. AIA-DAGA*, Meran, 2013.
- [13] V. Hansen and E. R. Madsen, "On aural phase detection," *J. Audio Eng. Soc.*, vol. 22, no. 1, pp. 10–14, January/February 1974.
- [14] —, "On aural phase detection: Part ii," *J. Audio Eng. Soc.*, vol. 22, no. 10, pp. 783–788, December 1974.
- [15] J. Blauert and P. Laws, "Group delay distortions in electroacoustical systems," *J. Acoust. Soc. Am.*, vol. 63, no. 5, pp. 1478–1483, 1978.
- [16] H. Suzuki, S. Morita, and T. Shindo, "On the perception of phase distortions," *J. Audio Eng. Soc.*, vol. 28, no. 9, pp. 570–574, September 1980.
- [17] S. P. Lipshitz, M. Pocock, and J. Vanderkooy, "On the audibility of midrange phase distortion in audio systems," *J. Audio Eng. Soc.*, vol. 30, no. 9, pp. 580–595, September 1982.
- [18] H. Møller, P. Minnaar, S. K. Olesen, F. Christensen, and J. Plogsties, "On the audibility of all-pass phase in electroacoustical transfer functions," *J. Audio Eng. Soc.*, vol. 55, no. 3, pp. 115–134, March 2007.
- [19] A. W. Lohmann, D. Mendlovic, and Z. Zalevsky, "Fractional Hilbert transform," *Opt. Lett.*, vol. 21, no. 4, pp. 281–283, 1996.
- [20] S. L. Hahn, *Hilbert Transforms in Signal Processing*. Boston: Artech House, 1996.
- [21] A. V. Oppenheim, R. W. Schaffer, and J. R. Buck, *Discrete-Time Signal Processing*. Upper Saddle River: Prentice Hall, 1999.
- [22] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*. New York: Dover, 1970.
- [23] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webMUSHRA—a comprehensive framework for web-based listening tests," *J. Open Res. Softw.*, vol. 6, no. 1, p. 8, 2018.
- [24] Recommendation ITU-R BS.1770-4, "Algorithms to measure audio programme loudness and true-peak audio level," ITU, Tech. Rep., 2017.
- [25] D. C. Howell, *Statistical Methods for Psychology*, 8th ed. Belmont: Wadsworth, Cengage Learning, 2013.
- [26] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behavior Research Methods*, vol. 38, no. 2, pp. 175–191, 2007.

<sup>4</sup> <https://github.com/spatialaudio/audibility-constant-phase>



## Full Reviewed Paper at ICSA 2019

Presented \* by VDT.

### Perceptual Evaluation of Spatial Resolution in Directivity Patterns 2: coincident source/listener positions

Matthias Frank<sup>1</sup>, Manuel Brandner<sup>1</sup>

<sup>1</sup> *Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Austria*  
*Email: frank@iem.at, brandner@iem.at*

#### Abstract

The incorporation of source directivity is important for a plausible and authentic auralization. While high-resolution measurement setups and data exist, it is yet not clear how detailed the directivity information has to be measured and reproduced with regard to perception. In particular, when source and listener are at the same location, resulting in a high direct-to-reverberant energy ratio, the precise shape of the directivity pattern might not yield perceptual differences. The paper approaches this question by a listening experiment in a virtual environment with generic directivity patterns and coincident position of listener and source. The experiment compares different spatial resolutions (spherical harmonic orders) of the directivity patterns for multiple virtual listener/source positions/orientations and levels of direct sound for speech and noise. The virtual environment employs a higher-order image-source model and binaural, dynamic Ambisonic playback. The results show that the exact shape of the directivity pattern is often perceptually irrelevant, while the preservation of the direct-to-reverberant energy ratio is more important.

#### 1. Introduction

Plausible and authentic auralization of sound sources in rooms benefits from the incorporation of source directivity and variable source orientation [1]. This is mainly due to the natural perception of distance that is controlled by the direct-to-reverberant energy ratio (DRR) [2, 3]. High-resolution measurement of source directivity is typically done with surrounding microphone arrays of up to 64 microphones at the same time [4] and directivity patterns are often represented in spherical harmonics to facilitate simple rotation. A high resolution is sometimes necessary to compensate for imprecise centering [5, 6], even for sources with low spatial resolution in their directivity patterns. Our previous study [7] revealed that perception of spatial resolution in directivity patterns is limited to spherical harmonic orders around 4 for large distances between source and receiver in a stimulated concert hall. In such cases, the DRR is typically negative.

However, for the auralization of one's own voice or when playing an instrument oneself [8–10], direct sound dominates.

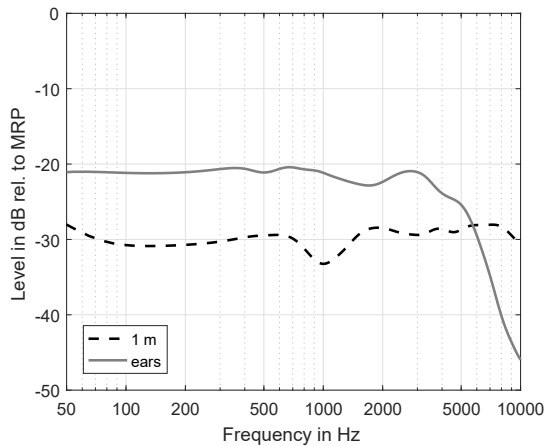
This paper investigates how precise directivity patterns are perceived in such cases, i.e. to which order a higher-order directivity pattern can be reduced to still be perceptually indistinguishable from a reference. The reference directivity pattern is highly directive as it appears for large brass instruments at high frequencies. The investigation employs a high level of direct sound as it appears in human speaking/singing and further, reduced levels to represent instruments with less direct sound at the player's ears. The virtual room in which the directional source is playing is simulated by an image-source model without late diffuse reverberation. These settings are chosen to simulate the most sensitive case, whereas a practical application might be less critical.

The paper first introduces setup and conditions of the listening experiment. The following section presents the experimental results. The results are subsequently compared to technical measures that are related to room acoustics and properties of the directivity patterns. Finally, the investigation is summarized and compared to our previous results in [7] for non-coincident listener and source.

## 2. Setup and Conditions

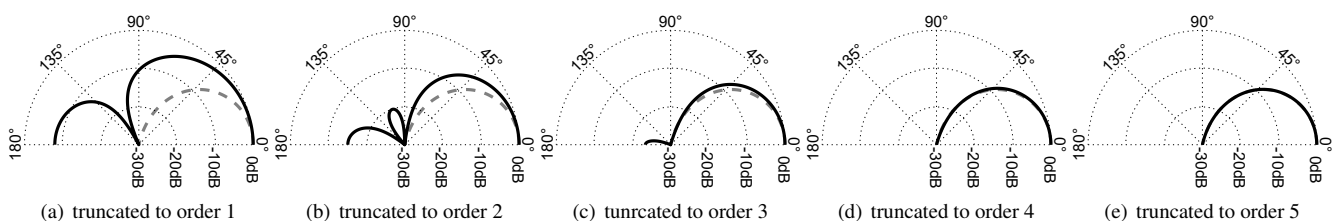
The parameters of the room simulation were identical to those used in [7]: The room had a size of 30 m × 20 m × 10 m and a reverberation time of 1.9 s between 200 Hz and 2 kHz, and was doubled/halved for frequencies below 100 Hz and above 4 kHz, respectively. The simulation employed a 7<sup>th</sup>-order image-source model (236 reflections) implemented in the IEM RoomEncoder VST plug-in<sup>1</sup>. The headphone playback employed 7<sup>th</sup>-order head-tracked [11] binaural Ambisonics [12] using the IEM BinauralDecoder. Note that the rotation of the source was linked to the head rotation.

The direct sound was not generated by the RoomEncoder plug-in, as this would result in an infinitely high level for coincident source and receiver position. Therefore, it was realized as omnidirectional sound inside the listener’s head with a specific level that should correspond to direct sound level at a speaker’s own ears. The level is based on a measurement of a B&K HATS 4128 using its mouth simulator, its ears, and two omnidirectional microphones at 1 m and 25 mm distance (mouth reference point, MRP) from the mouth in an anechoic chamber, respectively. Fig. 2 shows that the level at the ears is roughly 20 dB less than at the MRP. These results are similar to findings in [8] and the deviations can be explained by different distances of the MRP. The level in 1 m distance is again about 10 dB less than at the ears. A broad-band level difference of 10 dB was used to calibrate the direct sound and the image-source model for the experiment and is denoted as 0 dB direct sound level in the remainder of this paper. In order to represent instruments with less direct sound at the player’s ears, reduced levels { -10, -20 } dB were also evaluated.

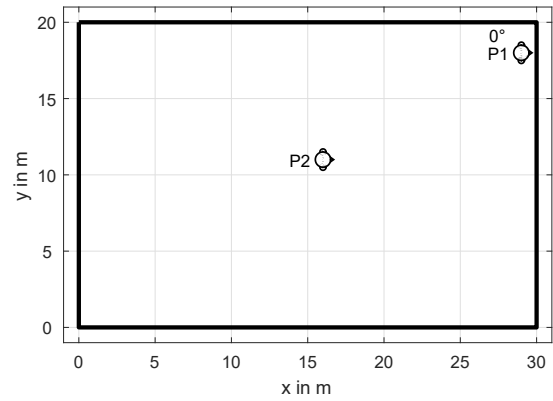


**Fig. 2:** Sound pressure levels in 1 m in front of the HATS and at its ears relative to the mouth reference point (MRP).

<sup>1</sup>freely available at plugins.iem.at



**Fig. 1:** On-axis equalized directivity patterns of beams in the experiment; gray dashed line indicates 7<sup>th</sup>-order inphase beam as reference.



**Fig. 3:** Listener/source position in the horizontal cross section of the simulated room. Indicated listener/source orientation is defined as 0°.

The source and the listener were positioned coincidentally with a height of 4 m above the floor at positions P1 and P2. P1 was close to a wall to provoke a strong first reflection that could interfere with the direct sound, when the source/listener was facing the wall (0° orientation), cf. Fig. 3. In contrast, the second orientation (180°) at P1 yielded weaker reflections. For P2, which was close to the center of the room, the reflection pattern was less orientation-dependent. Thus, there was only one orientation evaluated at P2.

The reference directivity was a 7<sup>th</sup>-order inphase [13] design, resulting in no side lobes and a relatively narrow main lobe, cf. Fig. 1. This directivity is similar to that of larger brass instruments, e.g. trombones or tubas, at high frequencies [14]. Typical directivity patterns of other instruments can be assumed to be less directive. In the experiment, the reference directivity pattern was reduced to orders 0 to 5 by simple truncation, as our previous study [7] revealed truncation to be perceptually better than preservation of nulls. Orders higher than 5 were excluded, as they were perceived as identical to the reference in preliminary tests. All resulting directivity patterns were diffuse-field equalized. The experiment employed two different sounds: (a) continuous pink noise for maximum sensitivity to coloration and (b) male English speech [15] that facilitates better spatial perception and familiarity.

Overall, there were 18 = 2 (sounds) × 3 ({0, -10, -20} dB direct sound level) × 3 (2 orientations at P1 + 1 orientation at P2) trials with multi-stimulus comparisons. The listeners task was to compare the similarity of the 6 (0<sup>th</sup> to 5<sup>th</sup> order truncation) stimuli to the corresponding 7<sup>th</sup>-order reference on a continuous scale from *very different* to *identical*. Note that the playback level in each trial was adjusted reversely to the level of the direct sound in order to achieve similar loudness between the trials.

### 3. Results

There were 10 experienced listeners (average age 31 years) who spent about 21 min each on the entire experiment. Based on the 10 values for each condition, the results of the experiment are presented as median values and corresponding confidence intervals in Figs. 4 and 5 for noise and speech, respectively. The gray level of the markers and lines in the figures indicates the level of the direct sound. Obviously, the similarity to the reference increases with the truncation order and also with the level of the direct sound for both sounds and all positions/orientations.

As we were interested in the spatial resolution required for perceptually indistinguishable auralization in comparison to the reference, Tab. 1 provides a suitable and easy-to-read representation of the results: For each condition, the table shows the minimum required order to yield indistinguishable results in terms of a Wilcoxon signed rank test with Bonferroni-Holm correction.

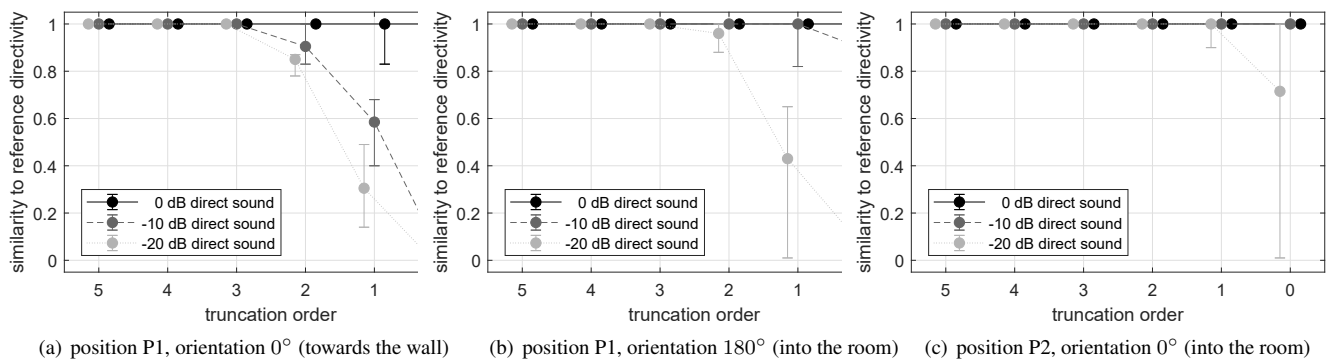
The influence of the direct sound can be seen clearly: While at the lowest level (-20 dB) orders around 2 are required, results are perceptually indistinguishable from the reference already

for an order of 0 at the highest level (0 dB) for all conditions except speech at P1 and 0° orientation. This indicates that for dominant direct sound, the exact control of the reflections by the directivity pattern is not important as long as the direct-to-reverberant energy ratio is preserved. This seems to be already assured by the diffuse-field equalization of the truncated directivity patterns.

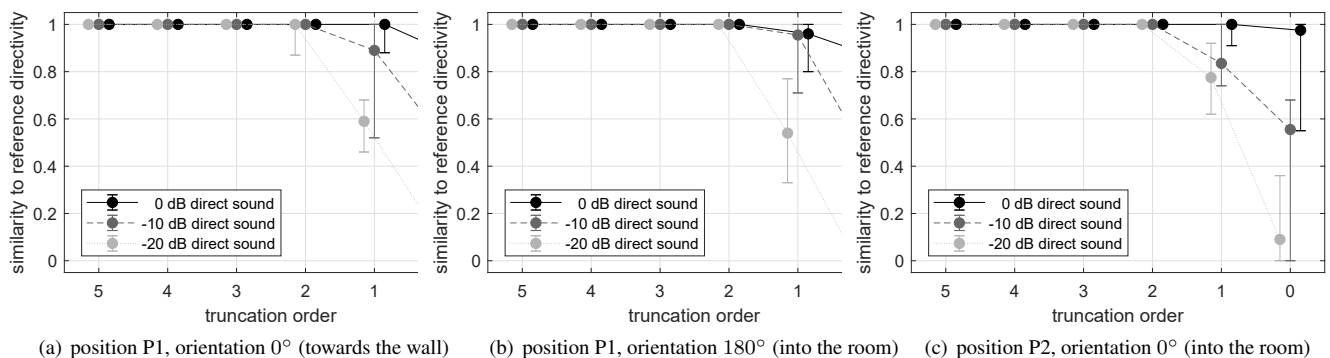
The sensitivity of the noise conditions increases with the proximity and orientation towards the walls: The central position P2 is most distant to all walls and it requires only an order of 1 or 0 for -20 dB or -10 dB direct sound, respectively. When facing the close wall at P1 and 0° orientation, orders of 3 and 2 are required for the same level of direct sound. In contrast, there is no dependency on the listener/source position and the orientation for speech, except for the increased sensitivity at P1 with 0° orientation. The increased sensitivity of noise in comparison to speech at P1 for 0° orientation and -20 dB direct sound is due to a strong comb filter. As listeners reported after the experiment, the truncation led to different strength of comb filters for noise, while it led to different level and density of reverberation for speech.

**Tab. 1:** Minimum required order to be indistinguishable from reference at 5% level with Bonferroni-Holm correction.

sound	0 dB direct sound			-10 dB direct sound			-20 dB direct sound		
	P1, 0°	P1, 180°	P2, 0°	P1, 0°	P1, 180°	P2, 0°	P1, 0°	P1, 180°	P2, 0°
noise	0	0	0	2	1	0	3	2	1
speech	1	0	0	1	1	1	2	2	2



**Fig. 4:** Medians and 95% confidence intervals of perceived similarity to auralization using 7<sup>th</sup>-order reference directivity for noise at different listening/source positions/ and orientations for different levels of direct sound.



**Fig. 5:** Medians and 95% confidence intervals of perceived similarity to auralization using 7<sup>th</sup>-order reference directivity for speech at different listening/source positions and orientations for different levels of direct sound.



### 4. Technical Measures

This section calculates some technical measures in order to generalize the experimental results for application on different room settings and directivity patterns.

The first kind of technical measure is the direct-to-reverberant energy ratio (DRR) and it depends on the combination of the directivity pattern, its orientation, the listener/source position, the direct sound level and the room. Note that in our calculation of DRR, the first reflections also contributed to the reverberant energy. Tab. 2 shows the resulting values in dependence of the direct sound level. Naturally, the DRR increases with the level of direct sound. The 0° orientation at P1 results in values about 16 dB lower than for the 180° orientation and P2 because it yields a strong first reflection from the nearby wall. In this case, the DRR increases for truncated orders due to a reduction of the reflection from the wall, i.e. the diffuse-field equalized directivity patterns radiate more energy into all other directions away from the wall. A similar, however weaker behavior can be seen at P2. In contrast, order truncation of the directivity pattern reduces the DRR for the 180° orientation at P1. Here, the lower-order patterns lead to an increase of the energy from the nearby wall that in turn reduces the DRR values.

Tab. 2 relates to the experimental results by printing values in bold that resulted in indistinguishable results for speech. For reference DRR values around 40 dB, deviations of around 4 dB were not perceivable. For values around 30 dB, deviations must not exceed 2 dB to remain perceptually irrelevant. Similar sensitivity can be found for DRR values around 0 dB. The tendency that sensitivity decreases towards higher DRR agrees with literature [16]. However, there are exceptions, where the threshold is smaller (P2, 0° with -20 dB direct sound: below 1 dB). This might be due to the different strategies for creating the stimuli: In [16], the direct sound was attenuated/boosted and the rest of the impulse response was kept identical. In our experiment, the modification of the directivity patterns modified the impulse response but the direct sound remained the same. In this way, we did not directly modify the level ratio between direct sound and reverberation, but the level of each reflection in the impulse response.

**Tab. 3:** Side lobes, beam width, and front-to-back energy ratio of the tested directivity patterns.

directivity order	side lobe in dB	width in °	F/B-R <sub>25</sub> in dB
7 (ref)	-∞	71	19.8
5	-49.1	72	19.8
4	-34.2	74	19.9
3	-23.4	81	20.1
2	-15.1	99	15.9
1	-8.0	147	9.5
0	0	360	0

The second kind of technical measures is independent of the room and the listener/source position because it solely depends on the directivity pattern itself. The measures are (a) side lobe: level of the strongest side lobe in dB, (b) width: aperture angle of the cap exceeding -6 dB relative to the maximum at the 0° direction in °, and (c) F/B-R<sub>25</sub>: front-to-back ratio in dB, with lower dynamic limitation at -25 dB relative to the maximum [7].

Tab. 3 shows the above-mentioned measures for the reference directivity and the directivities truncated at different orders. For -20 dB direct sound, the minimum required order for speech was 2. In this case, a side lobe attenuation of around 15 dB was not distinguished from the reference, a widening of the beam of 28° or 39%, and a F/B-R<sub>25</sub> difference of 3.9 dB. For noise under the most sensitive conditions, the required 3<sup>rd</sup> order resulted in a side lobe attenuation of around 23 dB, a widening of the beam of 10° or 14%, and a F/B-R<sub>25</sub> difference of 0.3 dB. Speech at -10 dB and all position/orientations, as well as at 0 dB at P1 with 0° orientation, required an order of 1, resulting in a side lobe attenuation of 8 dB, a widening of the beam of 76° or 107%, and a F/B-R<sub>25</sub> difference of around 10 dB. All other conditions with 0 dB direct sound did not require any modeling of the reference directivity except for diffuse-field equalization.

**Tab. 2:** Direct-to-reverberant energy ratio of the tested directivity patterns at the listener’s ears in dB for all listen/source positions and orientation. Values that resulted in indistinguishable results for speech are printed bold.

directivity order	0 dB direct sound			-10 dB direct sound			-20 dB direct sound		
	P1, 0°	P1, 180°	P2, 0°	P1, 0°	P1, 180°	P2, 0°	P1, 0°	P1, 180°	P2, 0°
7 (ref)	22.4	38.7	38.9	12.4	28.7	28.9	2.4	18.7	18.9
5	<b>22.5</b>	<b>38.7</b>	<b>38.9</b>	<b>12.5</b>	<b>28.7</b>	<b>28.9</b>	<b>2.5</b>	<b>18.7</b>	<b>18.9</b>
4	<b>22.7</b>	<b>38.7</b>	<b>38.9</b>	<b>12.7</b>	<b>28.7</b>	<b>28.9</b>	<b>2.7</b>	<b>18.7</b>	<b>18.9</b>
3	<b>22.9</b>	<b>37.9</b>	<b>38.7</b>	<b>12.9</b>	<b>27.9</b>	<b>28.7</b>	<b>2.9</b>	<b>17.9</b>	<b>18.7</b>
2	<b>24.5</b>	<b>36.1</b>	<b>39.0</b>	<b>14.5</b>	<b>26.1</b>	<b>29.0</b>	<b>4.5</b>	<b>16.1</b>	<b>19.0</b>
1	<b>27.8</b>	<b>33.9</b>	<b>40.8</b>	<b>17.8</b>	<b>23.9</b>	<b>30.8</b>	7.8	13.9	20.8
0	34.1	<b>34.1</b>	<b>42.4</b>	24.1	24.1	32.4	14.1	14.1	22.4

## 5. Conclusion

This paper evaluated the perceptual effect of reducing the spatial resolution (maximum spherical harmonics order) in directivity patterns for coincident source and listener position in a virtual room. For maximum sensitivity, the room simulation employed a higher-order image-source model without late diffuse reverberation and used dynamic binaural playback including head tracking that also controlled the orientation of the source. For the same reason, the reference directivity pattern was highly directive and the level of the direct sound was high, such as in human speech. The direct sound was played back omnidirectional, i.e. inside the listener's head and the evaluation also included conditions with reduced direct sound to simulate other instruments.

In comparison to our previous experiment [7] with non-coincident listener/source positions, the perceptual influence of the reduction in spatial resolution was less pronounced, i.e. lower spherical harmonic orders were required to produce perceptually indistinguishable results from the reference. This could be attributed to the dominance of the direct sound in the new experiment. Thereby, reducing the direct sound increased the minimum required orders from 0 to 2, on average. This result agrees with the literature [16], where the sensitivity of the direct-to-reverberant energy ratio (DRR) is highest for values around 0 dB and decreases towards large absolute values of the DRR. Although the reduction of the spatial resolution yields an increase in beam width and reduction of side-lobe attenuation, the DRR is often well preserved, especially at the central listener/source position and direct sound levels as in human speech. In such cases, the diffuse-field equalization of the reduced-order directivity patterns might already be good enough. However, when facing a nearby wall and with less direct sound, the preservation of the directivity pattern is more important. The perceptual effect of the order reduction seems to be signal-dependent: coloration for noise, level and density of reverberation for speech.

## Acknowledgments

This work is supported by the project Augmented Practice-Room (1023), which is funded by the local government of Styria via Zukunftsfonds Steiermark (future fond of Styria). The authors thank all listeners for their participation in the experiments and the reviewers for their helpful comments.



## References

- [1] B. N. J. Postma, H. Demontis, and B. F. G. Katz, "Subjective Evaluation of Dynamic Voice Directivity for Auralizations," *Acta Acustica united with Acustica*, vol. 103, no. 2, pp. 181–184, Mar. 2017.
- [2] A. Kolarik, S. Cirstea, and S. Pardhan, "Discrimination of virtual auditory distance using level and direct-to-reverberant ratio cues," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 3395–3398, 2013.
- [3] F. Wendt, F. Zotter, M. Frank, and R. Höldrich, "Auditory Distance Control Using a Variable-Directivity Loudspeaker," *MDPI Applied Science*, vol. 7, no. 7, 2017.
- [4] F. Hohl, "Kugelmikrofonarray zur Abstrahlungsvermessung von Musikinstrumenten," Master's thesis, TU Graz, 2009.
- [5] D. Deboy, "Tangential Intensity Algorithm for Acoustic Centering," in *Fortschritte der Akustik, DAGA*, Düsseldorf, 2011.
- [6] I. B. Hagai, M. Pollow, M. Vorländer, and B. Rafaely, "Acoustic centering of sources measured by surrounding spherical microphone arrays," *Journal of the Acoustical Society of America (JASA)*, vol. 130, no. 4, 2011.
- [7] M. Frank and M. Brandner, "Perceptual Evaluation of Spatial Resolution in Directivity Patterns," in *Fortschritte der Akustik, DAGA*, Rostock, Mar. 2019.
- [8] C. Pörschmann, "One's Own Voice in Auditory Virtual Environments," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 378–388, 2001.
- [9] J. S. Brereton, D. T. Murphy, and D. M. Howard, "The Virtual Singing Studio: A loudspeaker-based room acoustics simulation for real-time musical performance," in *Proceedings of the Baltic Nordic Acoustics Meeting (BNAM2012)*, 2012, pp. 18–20.
- [10] J. M. Arend, T. Lübeck, and C. Pörschmann, "A Reactive Virtual Acoustic Environment for Interactive Immersive Audio," in *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*, Mar. 2019.
- [11] M. Romanov, P. Berghold, M. Frank, D. Rudrich, M. Zaunschirm, and F. Zotter, "Implementation and Evaluation of a Low-Cost Headtracker for Binaural Synthesis," in *Audio Engineering Society Convention 142*, May 2017.
- [12] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares," in *Fortschritte der Akustik - DAGA*, Munich, March 2018.
- [13] J. Daniel, "Représentation des champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia," Ph.D. dissertation, Université Paris 6, 2001.
- [14] J. Pätynen and T. Lokki, "Directivities of Symphony Orchestra Instruments," *Acta Acustica united with Acustica*, vol. 96, no. 1, pp. 138–167, 2010.
- [15] EBU, "EBU SQAM CD: Sound Quality Assessment Material recordings for subjective tests," 2008. [Online]. Available: <https://tech.ebu.ch/publications/sqamcd>
- [16] E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng, "On the minimum audible difference in direct-to-reverberant energy ratio," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 450–461, 2008.





# Abstract Reviewed Paper at ICSA 2019

Presented \* by VDT.

## Investigation on spatial auditory perception using non-uniform spatial distribution of binaural room impulse responses

Stephan Werner<sup>1</sup>, Florian Klein<sup>1</sup>, and Georg Götz<sup>2</sup>

<sup>1</sup> *Technische Universität Ilmenau, Germany, Email: {stephan.werner; florian.klein}@tu-ilmenau.de*

<sup>2</sup> *Aalto University, Finland, Email: georg.gotz@aalto.fi*

### Abstract

For spatial audio reproduction in the context of virtual and augmented reality, a position-dynamic binaural synthesis can be used to reproduce the ear signals for a moving listener. A set of binaural room impulse responses (BRIRs) is required for each possible position of the listener in the room. The required spatial resolution of the BRIR positions can be estimated by spatial auditory perception thresholds. If the resolution is too low, jumps in perception of direction and distance and coloration effects occur. This contribution presents an evaluation of spatial audio quality using different spatial resolutions of the position of the used BRIRs. The evaluation is performed with a moving listener. The test persons evaluate any abnormalities in the spatial audio quality. The result is a comparison of the quality and the spatial resolution of the various conditions used.

## 1. Introduction

Existing audio systems can reproduce spatial audio in a way that artificial and real audio objects are perceived as plausible audible events in a virtual and/or augmented environment [3]. An auditory illusion of a spatial acoustic environment can be created with the help of existing position-dynamic binaural synthesis systems [2]. The occurrence of such a plausible auditory illusion depends on an adequate technical realization and on several context dependent quality parameters like congruence between synthesized scene and the listening environment or individualization of the technical system for example.

This paper examines spatial auditory perception thresholds using a position-dynamic binaural synthesis. The binaural transfer functions are provided for discrete positions in the room. The local area in which one set of BRIRs is used is referred to as cell. The size of the cell directly influences the direction and distance errors of the reproduction caused by it. The discretization of the room is therefore determined by the

perceived minimum direction and distance change between the position of the listener and the source. This allows the creation of a perceptually motivated BRIR grid which needs less BRIRs than a uniform shaped grid.

## 2. Binaural Synthesis System

The reproduction of an audio object in a reverberant room can be realized by using BRIRs. The audio signal of the source is convolved with the BRIRs for the left and the right ear and for the current source-receiver position and head orientation of the listener. A change of position and/or head pose requires a new pair of BRIRs. The position and pose changes are continuously measured by a tracking system and made available to the BRIR selection. In this contribution, a QualiSys motion capturing system is used to track the horizontal orientation and the x-y coordinates of the listeners' head. Headphones are mostly used as playback devices. An open or extra-aural headphone additionally enables an acoustic recognition of the real environment. The headphones

must be equalized for correct reproduction of the binaural ear signals. For this contribution a KEMAR head and torso simulator is used for the BRIR recordings and an equalized Beyerdynamic DT-1990pro headphone is used for playback. The inverse of the headphone transfer function is calculated by a least-square method with minimum phase inversion [13].

The needed BRIRs can be calculated by room-acoustic simulations or by measurements of real sound sources in a real room. A comprehensive synthesis of an auditory scene with a variety of sound sources, room acoustics, and movements of sources and receiver requires a high number of BRIRs. A minimization of this number while simultaneously maintaining a high perceived quality is desirable here.

Several approaches are available to reduce the amount of measurements for binaural synthesis. These are for example room acoustic simulations [12, 20] or methods which use head-related transfer functions and directional dependent or independent room impulse responses [6, 11]. In addition to these approaches the required BRIRs can be synthesized from the measurement of just one BRIR data set from only one position in the listening room by changing acoustic parameters like initial time delay gap, energy decay, and direct to reverberant energy ratio. These methods are applied for this contribution. Details can be found in the references [16, 18].

A basic feature of the used approach is the use of synthesized BRIR data sets for discrete positions in the room. These discrete areas can either have a uniform distribution (grids with rectangular grid areas/cells) or a non-uniform distribution of the single grid cells. Figure 1 illustrates an uniform grid. The listener can move within an area of max. 4 m x 4 m. The resolution of the binaural synthesis for translation is 0.25 m using an uniform grid of squares.

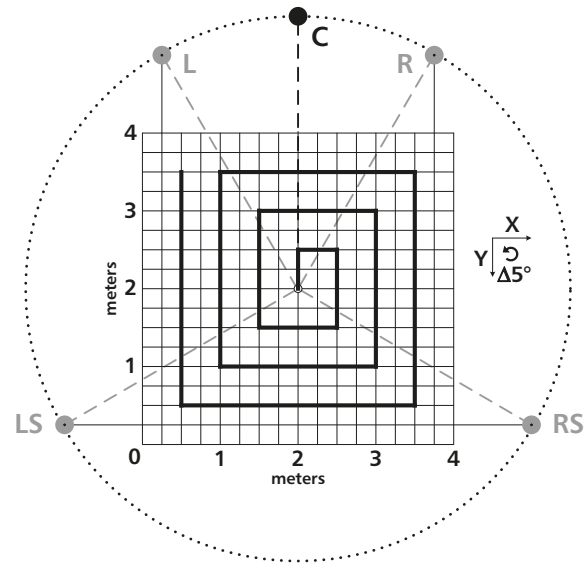
### 2.1. Basic Arrangement

Figure 1 illustrates the walkable area and the location of the possible audio object positions. The center speaker (C) is positioned at 0° and 3.5 m from the midpoint of an assumed circle. A left and a right speaker (L, R) is positioned at +/- 30°, another left and right speaker (SL, SR) is placed at +/- 120°. For the recording of the BRIRs, loudspeakers of the type Geithain MO2 were used. They were arranged at ear level of the used KEMAR artificial head. Their orientation was towards the midpoint of the area. Only the middle position of the area was measured. The other grid positions are synthesized as described above.

The concentric black line in figure 1 shows the path to be followed in the listening test by the listeners. More details about this can be found in section 3.2 Test Procedure.

### 2.2. Non-Uniform Grids

The use of a uniform grid as shown in figure 1 does not take perceptive inaccuracies in localization [7] and distance perception [1, 14] into account. Figure 2 shows the principle of localization blur and distance blur in the horizontal plane. If a fixed localization blur or minimum audible angle (e.g. 5°) is assumed, the density of the cells (in the sense of the

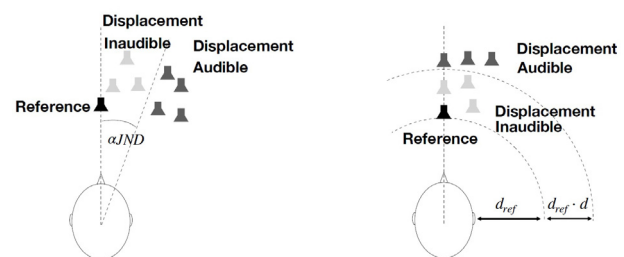


**Fig. 1:** Schematic view of the walkable area with the possible audio object positions. Concentric black line indicates the walking path.

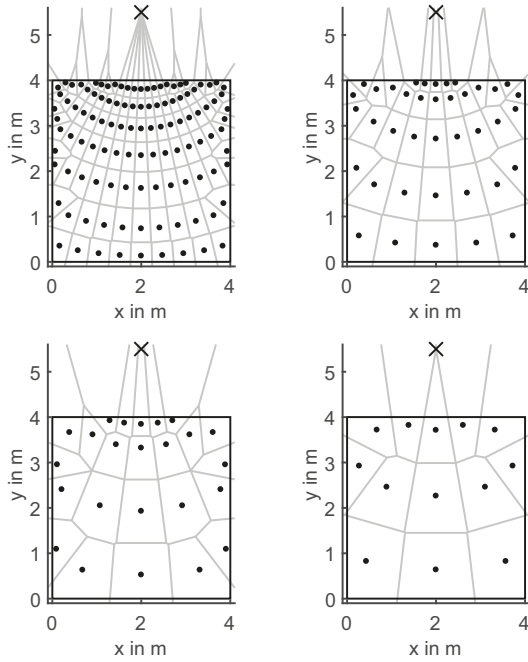
width of the cells) should have to increase at small distances to the source and decrease at larger distances. The situation is similar with regard to distance blur. For small distances the density (in the sense of the length of the cells) should be higher than for large distances. This approach leads to the non-uniform cells which can effectively save on the number of BRIRs required without causing increased errors in direction and distance perception.

In each individual grid cell, the BRIR set inside only represents the distance from the sound source to the center of the cell. If the listener is not in the center of the cell, the distance and direction cues no longer corresponds to the intended object position. The transition between the cells leads to a maximum deviation.

Figure 3 shows the four non-uniform grids used on the basis of a Voronoi network. Grid 1 (Figure 3 top left) is the finest grid and Grid 4 (Figure 3 bottom right) has the lowest resolution. The main parameters for generating the grids are the maximum allowed angle error  $\alpha$  and the maximum allowed relative distance change  $d$ . For the grids were used:



**Fig. 2:** Localization blur (left) and distance blur (right) in the horizontal plane. A displacement of the reference (black) by an angle or distance below the threshold is inaudible (light grey), whereas a bigger displacement is audible (dark grey). (from [4])



**Fig. 3:** Non-uniform grids used in the listening test; top left... grid 1, top right... grid 2, bottom left... grid 3, bottom right... grid 4.

Grid 1  $\alpha = 5^\circ$ ,  $d = 0.25$ ; Grid 2  $\alpha = 10^\circ$ ,  $d = 0.5$ ; Grid 3  $\alpha = 15^\circ$ ,  $d = 0.75$ ; Grid 4  $\alpha = 20^\circ$ ,  $d = 1.0$ . The distance parameter for Grid 1 is based on results from Spagnol et al. [15] where distance blurs of  $d = 0.25$  were found. The angle parameters are estimates which have been collected from the literature. It is assumed that the localization blur range is from  $1^\circ$  to  $10^\circ$ . The  $\alpha = 5^\circ$  for Grid 1 is selected on the one hand because it is in the middle of the indicated range and on the other hand it corresponds to the angular resolution of the BRIRs for head rotation. The perceptive effect of a small angle for localization blur would be masked in an evaluation by the larger angle for the head rotation.

For a detailed description of the procedure for creating non-uniform grids, we refer to the work of Georg Götz and Samaneh Kamandi [4]. This work can be requested from the authors of this contribution.

For each cell, a BRIR data set is stored for the left and right ear, which has a horizontal resolution of  $5^\circ$  for the head rotation. Therefore  $2 \times 72$  BRIRs are stored per cell and for each sound source. Table 1 shows the number of cells and number of BRIRs for the different grids. Compared to an assumed uniform grid with the same angle resolution of  $5^\circ$ , a grid resolution of 0.25 m, and for one source, as shown in Figure 1, there are high reductions in the number of BRIRs required. The reduction as a ratio of the number of BRIRs required for the uniform grid to the number of the sparsest non-uniform grid is 19.7 (see Table 1).

The centers of the individual cells are marked by dots in figure 3. Only at these points the BRIRs reproduce the correct direction and distance information. Outside the center, errors occur due to a shift in distance and direction. This effect is evaluated in the listening tests.

**Tab. 1:** Overview of the number of cells and BRIRs for the different grids. The ratio indicates the saving of BRIRs to the uniform grid.

Grid	number of cells	number of BRIRs	ratio
uniform	256	36,864	-
1	111	15,984	2.3
2	39	5,616	6.6
3	24	3,456	11.0
4	13	1,872	19.7

### 3. Listening Test

The listening test is intended to investigate various non-uniform grids with regard to spatial auditory perception. The aim is to find out what effect a reduction in the number of BRIRs has. Two evaluation approaches are used for this purpose.

#### 3.1. Test Design

The listening test is divided into two parts. In the first part, unanticipated events should be detected while the listener is walking around. These events should refer to artifacts in the spatial listening which may lead to an implausible scene perception. This can be for example: object extension, distance, localization, envelopment, externalization, spaciousness, and timbre.

The second part of the test takes place each time a test condition, and therefore a path, has been completely passed through. The test person evaluates various quality features in a questionnaire (see subsection Quality Evaluation 3.1.2). Of course, this does not make the single unanticipated event assessable, but at least the quality impression of the single test conditions.

##### 3.1.1. Detection of Unanticipated Events

The test persons should confirm a detected unanticipated event while walking around with the help of a radio button. The time and especially the x-y coordinate in the grid is recorded as information using the tracking system for the binaural synthesis. The button itself is realized by a presenter for slide presentations via a radio link and queried with the help of a Python script.

The test person is allowed to explore the local area more precisely by small forward and backward movements and head rotations when an unanticipated event is detected. This should allow a higher reliability in the determination of the position.

##### 3.1.2. Quality Evaluation

The evaluation of perception is performed for different quality features. In addition to single features, this also includes the rating of the overall impression as a kind of overall quality. The individual quality features include localization, externalization, and timbre perception. The query usually takes place on quasi-continuous scales with the negative value at "0" and the positive value at "100". Externalization is rated on a category scale.



**Overall Impression OI** - The overall impression should capture the individual impression and the perception of the overall quality of what is heard by the listeners. No specifications were made with regard to possible underlying single quality features. The survey of the overall impression was always carried out at the beginning of the survey for each walking path in order to minimize the influence of the evaluation of single quality features. The following question had to be answered: "How would you rate the quality of experience of this system?" A quasi-continuous rating scale from "0-poor" to "100-very good" was used.

**Localization Ability LA** - The localization ability is intended to test the ability of the listener to localize the auditory event. A high localization ability exists when the auditory event can be clearly assigned to a direction. If the directional information of the auditory event is diffuse or not localizable, a low rating should be given. The following question had to be answered: "How would you rate the ability to localize the audio object?" The quasi-continuous scale used ranges from "0-poor" to "100-very good".

**Localization Stability LS** - The use of different grids with different spatial resolution can lead to jumps in the perception of the object position. The movement of the listener through the individual grid cells causes more or less pronounced abrupt changes in the reproduction of directional and distance cues. The following question had to be answered: "How would you rate that the audio object stays at a fixed position?" The quasi-continuous scale used ranges from "0-poor" to "100-very good".

**Coloration CO** - The feature coloration aims at the evaluation of the hearing perception, which cannot be described by directional hearing, perception of externalization or spatial stability of the auditory event. The underlying perceptual feature of coloration is timbre. Timbre can be defined as "that attribute of sensation in terms of which a listener can judge that two steady complex tones having the same loudness, pitch and duration are dissimilar" [9]. In the present case, timbre is defined as the difference in perception of the audio signal at a detected abnormality and the otherwise perceived audio. This is referred to as coloration. The audio signal of the abnormality is evaluated as colored compared to the remaining audio shortly before it. The following question had to be answered: "How would you rate the coloration of the audio during walking?" The quasi-continuous scale is in a range from "0-strong coloration" to "100-no coloration".

**Externalization** - The externalization of auditory events describes the perception of the location of the event in the head or outside the head of the listener [5, 10]. Externalization is a crucial feature to reach a plausible spatial auditory illusion with binaural headphone systems [10, 17]. The dichotomous quality feature is counted as the index of the ratings on a three-point scale. In addition to the characteristics "in-head" and "outside the head" a transition point "outside but close to the head" is used. This scale is motivated by the individual mapping to a scale of the percept of externalization for every

test person. Only the scale point "outside the head" is counted as an externalized auditory event in further analysis. We define the perception of an event very close to the head or ears as in-head-localized or non-externalized. We suppose that this conservative approach maps the ratings in a reliable way referred to the resynthesis of the real loudspeakers with their positions in the room. The goal is to minimize the confusion between distance perception and externalization for closer distances. The following questions had to be answered: "How would you rate the externalization of the audio object?". The following scale is used "1=in-head", "2=close to the head", "3=in the room", "4=at the intended distance".

### 3.2. Test Procedure

The test was divided into several phases. The first phase comprises a written and oral introduction to the test environment, the assessment methodology and the quality characteristics to be assessed. The second phase is the familiarization phase with the position-dynamic binaural synthesis system. The headphones were placed on the listeners and were not removed until the end of the entire test. This should make it possible to get used to the headphones. The test persons should continue to move freely within the accessible area. Two audio scenes were synthesized binaurally.

The **first phase** consists of different male and female speakers who are placed at the five loudspeaker positions. The scene starts with one speaker until all five speakers are active at the same time. The intention of this scene is to capture individual sound object positions, to get used to a complex scene and to promote the active movement of the listener by specifically listening to individual sound sources, also by moving towards the sources.

The **second phase** is an excerpt from a radio play in which there are discrete audio objects and an enveloping and ambient sound-scape. In contrast to the first scene, this scenery is realized by multi-channel stereo panning onto the five loudspeakers. The intention of this scene is to get used to a complex environment and to listen to the content without consciously considering the technical realization. A uniform grid with a spatial resolution (spacing of the grid cells) of 0.25 m and an angular resolution of the horizontal head rotation of 5° was used for playback. This resolution leads to slight perceptible artifacts in critical hearing in the form of localization jumps of the audio object position in translational and rotational movements. A technical and perceptive proof of function of such a system is to be taken from the references [8, 18, 19].

The **third and final phase** includes the evaluation of the individual grid conditions. At the beginning, a section of an audio book is presented to the test persons. The intended source position is the center position of the arrangement used. The test persons should follow the path shown in figure 1 in order to cover an area as large as possible uniformly. The concentric path starts in the middle of the walkable area and leads to the outside. Its length is 24m for one walking direction. The path had to be walked forward and backward. According to the instructions the test persons had to mark

unanticipated events by pressing a radio button. At first the worst resolution grid was presented. This should ensure that abnormalities are also found. The different test conditions (different grids and audio signals) are then presented in random order for each test person. After the path for each condition had expired, the individual quality characteristics were evaluated in a questionnaire. After this the next test condition is presented. The test persons were instructed to position their heads in the walking direction. However, head movement was explicitly allowed. The test persons were also instructed to move in a normal to slow walking speed. After the evaluation of the last test condition, the test persons had the opportunity to make further comments and remarks on the test. The whole procedure took about one hour.

### 4. Ratings

The evaluations of 21 out of 22 test persons with a mean age of 29.1 years (standard deviation 8.2 years) were used for the evaluation. Fifteen of the five women and 17 men were experienced in listening tests and ten persons had experiences with binaural synthesis systems. Experience is defined as participation in at least two listening tests or the listening to a binaural synthesis at least twice. The ratings of one test person were omitted because even for the playback with the worst grid, ratings were always given on the scales of more than 85 scale points and thus clearly outside the average of the ratings of the other test persons. No further post-screening of the assessments was carried out. The authors are not aware of any method that allows a reliable and comprehensible evaluation of spatial hearing quality assessments using a virtual hearing instrument. The test itself did not contained any evaluations of real sound sources that might have allowed this.

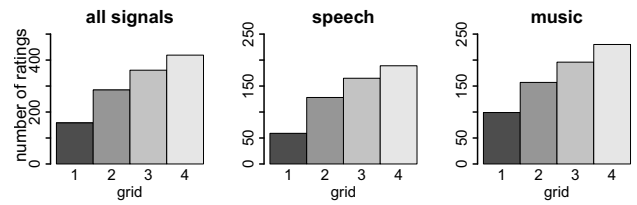
Table 2 shows the time required for the path to walk forward and backward. The average speed corresponds to a slow walking speed. The standard deviation as well as the max and min values indicate a certain variance in walking speed between the test persons. The mean walking speed was at 0.4m/s, with a length of the whole path of 48m.

**Tab. 2:** Duration of walking of the test persons to complete the path (forward and backward); times are in minutes; N indicates the number of the single walks.

N	mean	standard deviation	max	min
160	02:01	00:22	03:02	01:17

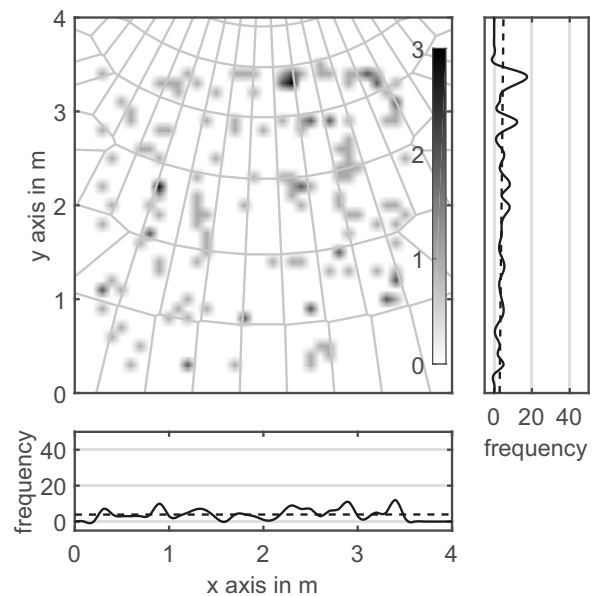
#### 4.1. Detection of Unanticipated Events

Figure 4 shows the total number of ratings for the detection of unanticipated events during listening. The number is proportional to the grid resolution. There is an almost linear increase with relation to the grids, which indicates an uniform selection of the grids with relation to its parameters. There is a higher number of ratings for the music signal than for the speech signal. This may indicate that the music signal was rated as a more critical signal. The test subjects' statements at the end of the test also indicate this.



**Fig. 4:** Absolute number of ratings of unanticipated events for all grids and audio signals.

The figures 5 and 6 show the frequency of perceived unanticipated events as a heat map while walking around. The side plots show the sum of the frequencies for the x and y axis as histograms. The gray lines shown in the heat map illustrate the position of the respective grid cells. Only the ratings for the most highly resolved grid (grid 1, figure 5) and the least resolved grid (grid 4, figure 6) are shown. The ratings for the other two grids lie between these two.

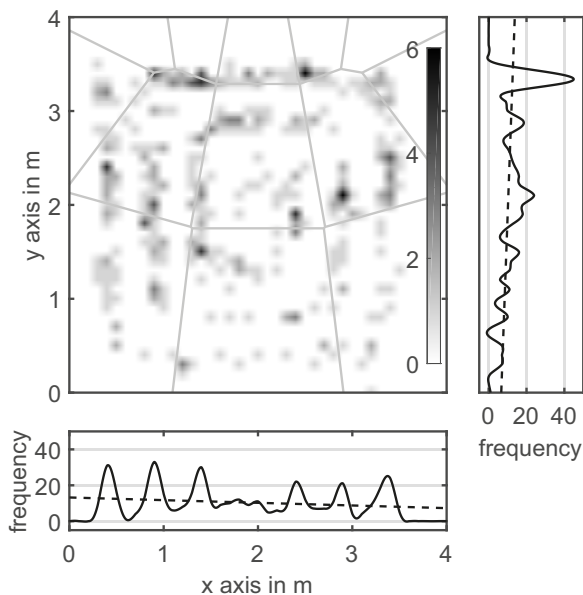


**Fig. 5:** Ratings of unanticipated events as heat-map for Grid1 ( $\alpha = 5^\circ; d = 0.25$ ). Side plots show the sum of the ratings over the x or y axis as a histogram. Dashed lines indicate regression line. Grey lines illustrate the position of the respective grid cells.

When the audio signals are played back using Grid 1 (Figure 5), the number of detected abnormalities is significantly lower compared to grid 4 (Figure 6). For all grids there is tendentially a symmetrical distribution of the abnormalities along the x-axis. For the y-axis there is a decrease in the number of abnormalities with increasing distance to the source. This effect increases with the use of more coarse grids.

#### 4.2. Quality Evaluation

Figure 7 shows the ratings of perceived externalization of auditory events for the various grids and audio signals. When using the grid with the highest resolution (Grid 1), high externalization indices close to 1 are visible. The indices decrease continuously but slightly to a value of around 0.75

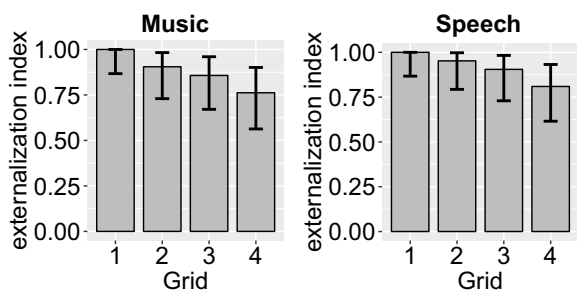


**Fig. 6:** Ratings of unanticipated events as heat-map for grid 4 ( $\alpha = 20^\circ; d = 1.0$ ). Side plots show the sum of the ratings over the x or y axis as a histogram. Dashed lines indicate regression line. Grey lines illustrate the position of the respective grid cells.

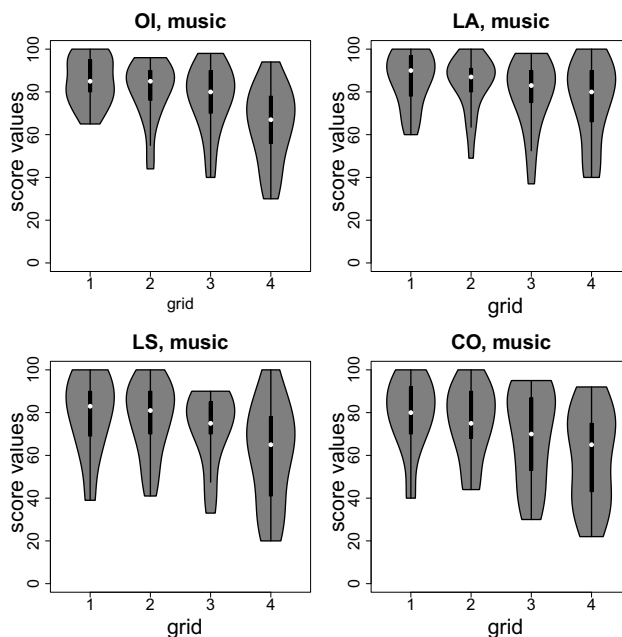
when less high-resolution grids are used. It is assumed that the localization and distance errors lead to a reduction of the externalization due to a reduced plausibility of the perception. With regard to the audio signals used, there are no influences on the externalization observed.

Figures 8 and 9 show the ratings for the individual quality features and for the overall impression for the music and the speech signal. It can be said that Grid 1 achieved the highest ratings. There is a tendency for lower valuations to be given for the less resolved Grids. The interquartile distances tend to rise slightly for Grid 3 and 4. This indicates that the test persons are less in agreement if grids with lower resolutions are used.

Overall it must be noted that relatively high quality ratings are given both for the overall impression and for the individual quality features in view of the savings achieved in the number of required BRIRs. However, it must also be made clear that the perceived quality also depends on the type of application. For example, a test scenario with a direct comparison of real and virtual sources with the intention of testing for



**Fig. 7:** Externalization as index with 95% conf. interval; left: music, right: speech.

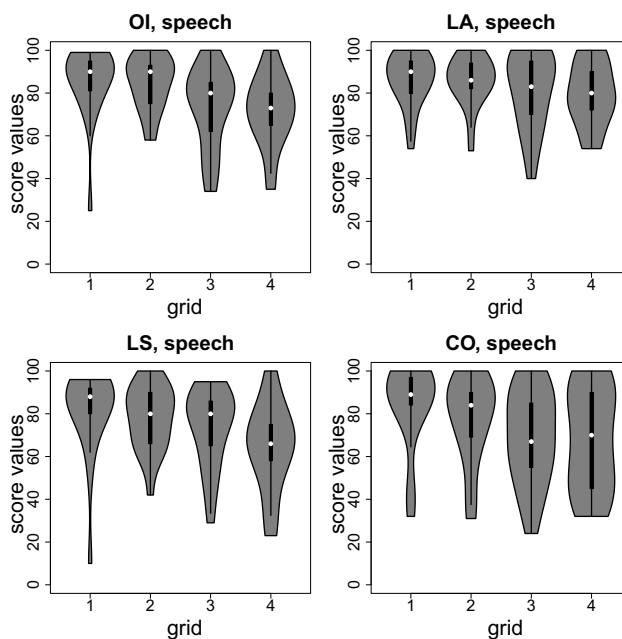


**Fig. 8:** Ratings of the quality features for the different grids and for the *music signal* as violin plots including boxplots; OI... Overall Impression, LA... Localization Ability, LS... Localization Stability, CO... Coloration.

authenticity would most likely lead to a much more critical evaluation [3].

### 5. Conclusion

The evaluation with regard to the detection of abnormalities and on the quality assessment shows that the test persons are



**Fig. 9:** Ratings of the quality features for the different grids and for the *speech signal* as violin plots including boxplots; OI... Overall Impression, LA... Localization Ability, LS... Localization Stability, CO... Coloration.

quite capable of detecting abnormalities in the various grid conditions. The low-resolution grids are rated worse than the high-resolution grids. However, this effect is much less pronounced for the individual quality features and for the overall impression. It seems to be the case that although the test persons recognize abnormalities, these only lead to a slight decrease of the perceived quality and plausibility of the scene. A further evaluation of authenticity and plausibility can provide information on this. This will also include more critical test signals (e.g. noise), a comparison between uniform grids and quasi-continuous provision of BRIRs. This will be done at a later date.

For a further evaluation in this field, it is intended to consider the assessments for the individual test persons. It can already be seen that the test persons are not homogeneous in their assessments. There are people who generally rate more critically than others.

Nevertheless, the results show that the use of non-uniform grids leads to a suitable position-dynamic binaural synthesis. The number of BRIRs used can be significantly reduced while maintaining a plausible spatial auditory perception. The presented method of non-uniform grids can be applied to any local provision of BRIRs. Furthermore, the method can be used to determine perception differences and perception thresholds in relation to directional and distance perception with a moving listener.

## 6. Acknowledgment

For their participation in the listening tests and for their interest in research, we would like to thank the test persons. Furthermore, we thank the students of the lecture "Advanced Psychoacoustic" of the Media Technology course at the TU Ilmenau. Special thanks go to Georg Götz and Samaneh Kamandi for working on their media project [4] in our group.

## 7. References

- [1] Daniel H. Ashmead, Deford Leroy, and Richard D. Odom. 1990. Perception of the relative distances of nearby sound sources. *Perception and Psychophysics* 4, 47 (1990), 326–331.
- [2] Karlheinz Brandenburg, Estefania Cano, Florian Klein, Thomas Koellmer, Hanna Lukashovich, Annika Neidhardt, Ulrike Sloma, and Stephan Werner. 2018. Plausible Augmentation of Auditory Scenes using Dynamic Binaural Synthesis for Personalized Auditory Realities. In *Audio Engineering Society Conference on Audio for Virtual and Augmented Reality, Redmond, USA*.
- [3] Fabian Brinkmann, Alexander Lindau, and Stefan Weinzierl. 2017. On the authenticity of individual dynamic binaural synthesis. *J. Acoust. Soc. Am.* 142, 4 (oct 2017), 1784–1795. <https://doi.org/10.1121/1.5005606>
- [4] Georg Goetz and Samaneh Kamandi. 2018. Optimization of the number and spatial distribution of binaural room impulse responses in an augmented auditory reality application. Media Project, Technische Universität Ilmenau, Electronic Media Technology Group.
- [5] William M. Hartmann and Andrew T. Wittenberg. 1996. On the externalization of sound images. *J. Acoust. Soc. Am.* 6, 99 (1996), 3678–3688. <https://doi.org/10.1121/1.414965>
- [6] Gavin Kearney, Claire Masterson, Stephen Adams, and Frank Boland. 2009. Towards Efficient Binaural Room Impulse Response Synthesis. In *EAA Symposium on Auralization, Espoo, Finland*.
- [7] A. W. Mills. 1958. On the minimum audible angle. *J. Acoust. Soc. Am.* 4, 30 (1958), 237–246. <https://doi.org/10.1121/1.1909553>
- [8] Christina Mittag, Stephan Werner, and Florian Klein. 2017. Development and Evaluation of Methods for the Synthesis of Binaural Room Impulse Responses based on Spatially Sparse Measurements in Real Rooms. In *43. Jahrestagung fuer Akustik, DAGA, Kiel, Germany*.
- [9] Brian C. J. Moore. 2012. *An introduction to the psychology of hearing*. 6th edition, Emerald Group Publishing Ltd, London, United Kingdom.
- [10] Georg Plenge. 1972. Ueber das Problem der Im-Kopf-Lokalisation. *Acustica* 26, 5 (1972), 241–252.
- [11] Christoph Poerschmann, P. Stade, and J.M. Arend. 2017. Binauralization of omnidirectional room impulse responses-algorithm and technical evaluation. In *20th Int. Conf. on Digital Audio Effects (DAFx), UK*. 345–352.
- [12] Lauri Savioja, Jyri Huopaniemi, Tapio Lokki, and Riitta Vaeaenaenen. 1999. Creating Interactive Virtual Acoustic Environment. *J. Audio Eng. Soc.* 47, 9 (1999), 675–705.
- [13] Zora Schaerer and Alexander Lindau. 2009. Evaluation of Equalization Methods for Binaural Signals. In *Audio Engineering Society Convention 126*. <http://www.aes.org/e-lib/browse.cfm?elib=14917>
- [14] W. E. Simpson and D. Stanton Lee. 1973. Head movement does not facilitate perception of the distance of a source of sound. *The American Journal of Psychology* 1, 86 (1973), 151–159.
- [15] Simone Spagnol, Rebekka Hoffmann, Arni Kristjansson, and Federico Avanzini. 2017. Effects of stimulus order on auditory distance discrimination of virtual nearby sound sources. *J. Acoust. Soc. Am.* 4, 141 (2017), 375–380.
- [16] Stephan Werner. 2018. *Ueber den Einfluss kontextabhängiger Qualitätsparameter auf die Wahrnehmung*

von *Externalitaet und Hoerereignisort [On the influence of context-dependent quality parameters on the perception of externality and auditory event location]*. Ph.D. Dissertation. Technische Universitaet Ilmenau, Faculty of Electrical Engineering and Information Technology, urn:nbn:de:gbv:ilm1-2018000672, Ilmenau, Germany.

- [17] Stephan Werner, Florian Klein, Thomas Mayenfels, and Karlheinz Brandenburg. 2016. A Summary on Acoustic Room Divergence and its Effect on Externalization of Auditory Events. In *8th International Conference on Quality of Multimedia Experience (QoMEX), Lisbon, Portugal*. <https://doi.org/10.1109/QoMEX.2016.7498973>
- [18] Stephan Werner, Annika Neidhardt, Florian Klein, and Karlheinz Brandenburg. 2018. Comparison of Different Methods to Create an Interactive Augmented Auditory Reality Scenario Using Sparse Binaural Room Impulse Response Measurements. In *44. Jahrestagung fuer Akustik, DAGA, Garching, Germany*.
- [19] Stephan Werner, Mina Voigt, Florian Klein, and Annika Neidhardt. 2018. A position-dynamic binaural synthesis of a multi-channel loudspeaker setup as an example of an auditory augmented reality application. In *30th Tonmeistertagung VDT International Convention, Cologne, Germany*.
- [20] Andreas Zimmermann and Andreas Lorenz. 2008. LISTEN: a user-adaptive audio-augmented museum guide. *User Modeling and User-Adapted Interaction* 18, 5 (2008), 389–416. <https://doi.org/10.1007/s11257-008-9049-x>





# Abstract Reviewed Paper at ICSA 2019

Presented \* by VDT.

## Real-time Estimation of Reverberation Time for Selection of suitable binaural room impulse responses

F. Klein<sup>1</sup>, A. Neidhardt<sup>1</sup>, M. Seipel<sup>1</sup>

<sup>1</sup> *Technische Universität Ilmenau, Germany, Email: florian.klein@tu-ilmenau.de*

### Abstract

The aim of auditory augmented reality is to create a highly immersive and plausible auditory illusion combining virtual audio objects and scenarios with the real acoustic surrounding. For this use case it is necessary to estimate the acoustics of the current room. A mismatch between real and simulated acoustics will easily be detected by the listener and will probably lead to In-head localization or an unrealistic acoustic envelopment of the virtual sound sources. This publication investigates State-of-the-Art algorithms for blind reverberation time estimation which are commonly used for speech enhancement algorithms or speech dereverberation and applies them to binaural ear signals. The outcome of these algorithms can be used to select the most appropriate room out of a room database for example. A room database could include pre-measured or simulated binaural room impulse responses which could directly be used to realize a binaural reproduction. First results show promising results combined with low computational effort. Further strategies for enhancing the used method are proposed in order to create a more precise reverberation time estimation.

## 1. Introduction

The aim of Auditory Augmented Realities (AAR) is to enrich the real acoustic environment of the listener with additional sound objects. Thus, it is inevitable to match the acoustic of the augmented sound objects with the acoustics of the real room or acoustic surrounding. Previous research has shown that a mismatch of the virtual and real acoustics can lead to in-head localization which means, that the sound objects are not located outside the head as usual for natural sound sources [12]. This effect was named room-divergence effect [9]. The current study focuses on the reverberation time as room dependent value, but of course a room has much more characteristics and properties which contribute to the “identity of a room”. The pattern of the first reflections, timbre of the reverb, room acoustical modes and others are maybe equally important. Additionally, many of these parameters change when the listener moves through the room which gives him or her an impression how the room sounds.

In order to realize a listening room dependent acoustic simu-

lation, it is possible to create a 3D model of room and conduct a simulation [11] or to select an appropriate set of binaural room impulse responses (BRIRs) out of a database. On the one hand, approaches which generate or measure BRIRs beforehand require an extensive database but on the other hand the computational load during the rendering of the ear signals is low, because only the BRIR selection has to be done in real time. In conjunction with methods for the interpolation and extrapolation of BRIRs of a moving listener [10], the required memory for a BRIR database can be reduced. This approach is the motivation behind this publication.

## 2. State of the Art

Auditory Augmented Realities is a emergent research field and there are no state of the art approaches to include the real acoustics into the simulation. But of course there are several ways to do so. A database of rooms and acoustic surroundings can be used in conjunction with artificial neural networks for



	Lecture Room	Meeting Room	Office Room	H2505
Size [m]	10.8 x 10.9 x 3.15	8 x 5 x 3.1	5 x 6.4 x 2.9	9.9 x 4.7 x 3.1
Meas. Dist. [m]	4, 5.56, 7.1, 8.68, 10.2	1.45, 1.7, 1.9, 2.25, 2.8	1, 2, 3	1, 2, 3, 4, 5, 6, 7, 8

**Tab. 1:** Dimensions of the rooms from which the BRIRs were used in this study. Lecture, meeting and the office rooms are located at the RWTH Aachen and the H2505 room at the TU Ilmenau. The measurement positions indicate the different distances between sound source and artificial head.

acoustic scene classification [14]. Those methods may also be able to estimate room acoustic parameters or room geometries in the future. By including optical information a room size or reverberation estimation might be improved. When multiple cameras or time of flight cameras are available, a 3D model of a room can be created [15]. Based on this, an acoustic room simulation can be realized which synthesizes new binaural room impulse responses. This paper focuses on blind T60 estimation methods known from communication engineering. Blind T60 estimations are utilized for dereverberation in order to enhance speech perception, for automatic speech recognition or for acoustical scene analysis [4]. Because of these popular applications several methods for blind reverberation time estimation from speech exist [6]. In a comparative study by Eaton [5] the algorithms by Prego [7] and Löllmann [8] delivered the best results. Both algorithms use framewise subband analysis of free decaying regions. Eaton states, that algorithms which use features based on the decay rate are most accurate.

### 3. Reference and test data

Two different datasets of binaural room impulse responses (BRIRs) were used for this work. A freely available data set from Jeub et al. [3] includes BRIR measurements of three different rooms at the RWTH Aachen University. Another data set was recorded at the TU Ilmenau in room H2505, which is a seminar room. Thus, four different rooms with different dimensions and acoustical properties were available (see table 1 for details). In each room, measurements with different distances to the sound source were conducted. For the measurements in the lecture, meeting and office room the artificial head HMS2 from HEAD Acoustics was used. For BRIR measurements in Aachen, pseudo-random noise was used to excite the space [3]. The measurements in room H2505 were conducted with an artificial head type KEMAR 45BA of the manufacturer G.R.A.S. The room H2505 was excited with a sweep from 60Hz to 20kHz. For all positions BRIRs with head direction of  $0^\circ$  azimuth were used. The default orientation of the sound source is frontal towards the dummy head. For room H2505 there was an additional condition with the sound source turned to the opposite direction.

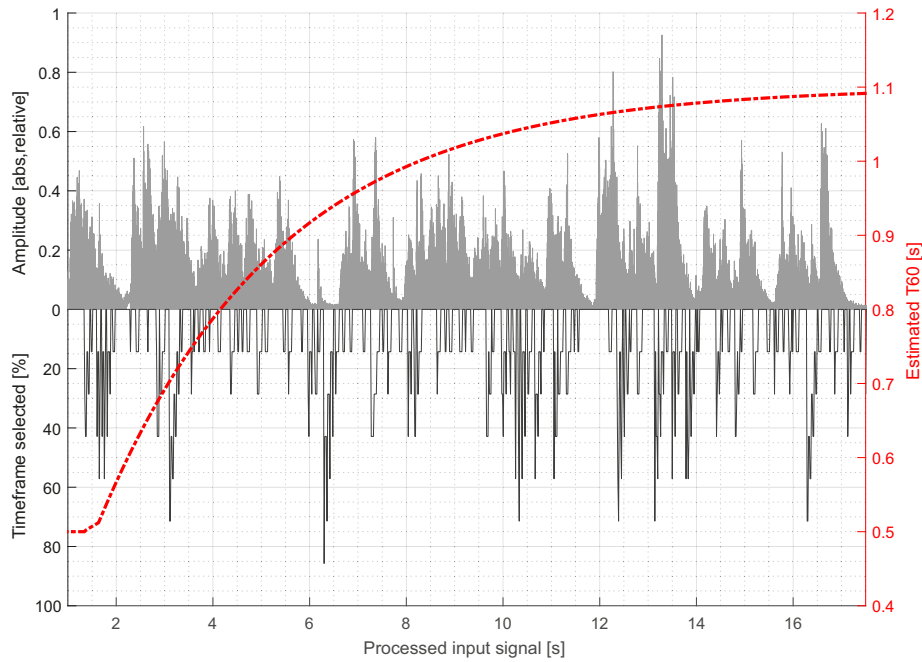
A reference T60 dataset was derived from these measurements. At first the T30 is calculated on basis of the energy decay curve of the BRIRs and then the T60 value is calculated from the T30 value. This calculation is done for left and right BRIR and the T60 valued are averaged afterwards. This is justifiable, because the head is oriented towards the source and therefore left and right ear signals should be similar.

However, there were small differences because the room acoustics were not symmetric or the measurement positions were not centered in the rooms.

For the preparation of the test data, the BRIRs are convolved with a 30s speech signal. Three additional variants were created by adding white noise to create a signal with 30dB, 20dB and 10dB SNR next to the basic version without artificially added noise. This was done to simulate more realistic and less sophisticated recording equipment like microphones typically used in consumer electronics.

### 4. T60 estimation algorithm

An algorithm proposed by Löllman et al. [2] was used to estimate the reverberation time. In contrast to the aforementioned algorithm Löllman (see section 2) this algorithms makes no use of a subband decomposition. The advantages are a low computational complexity and robustness against background noise [5]. The algorithm uses a statistical decay model of the transient signal components from a mono speech signal and their selection for a blockwise T60 calculation. The selection of interesting frames is done via a maximum value, minimum value, and energy comparison of successive subframes within every frame. The calculated values are compared with those of the previous subframe and it is determined whether the entire frame has a descending structure. If this is the case, the frame is considered in the T60 calculations. If this is not the case, the current frame is discarded and the next frame is analyzed. For each subframe a decay parameter is determined which can be used for T60 calculation. With the help of a maximum-likelihood estimation, the most probable reverberation time is determined by the probability density of the decay parameter of the selected frame, thereby calculating the T60 for each frame. The calculated values are stored in a T60 histogram. The variance of the stored T60 data is last reduced by a recursive smoothing to get a more reliable final estimation. The binsize of the histogram describes the quantization step in which the T60 values are calculated. For the calculations, the smoothing factor  $\alpha$  and the binsize were adapted to achieve the best possible results for the T60 values in this case. The settings of the smoothing factor and binsize have an effect on the accuracy of the estimation for different reverberation times. A smaller binsize and a smaller smoothing factor is beneficial for shorter reverberation times and respectively, the algorithm archives better results for longer reverberation times with a larger binsize and a higher smoothing factor. A higher downsampling factor speeds up the process. The settings for this study were determined empirically. Table 2 shows the selected values.



**Fig. 1:** Example of the frame wise processing of the input signal. Figure shows the waveform of a speech signal in room H2505, the percentage of each frame which was used for calculation as well as the progression of the T60 estimation.

frame size [n]	5740
subframe size [n]	820
binsize	0.15
smoothing factor $\alpha$	0.996
downsampling factor $D$	2
input size [s]	30

**Tab. 2:** Empirically selected parameters for the T60 estimation algorithm by Löllman.

The algorithm works in an interval of the reverberation time from  $0.2s$  to  $1.2s$  and with a recommended minimum input signal duration of  $10s$ . The quality of the estimate changes with the SNR of the input signal. The decaying structure of the transients is distorted by the noise components and fewer frames are considered during the preselection of frames, which leads to an inaccurate T60 estimate. Therefore, input signals with a low SNR need to be preprocessed with a noise reduction algorithm to ensure a good estimate. Noise reduction was performed using a wavelet denoising process. The denoising was designed adaptively for various SNR. For a  $SNR > 20dB$  no noise reduction is carried out, for  $20dB \leq SNR < 10dB$  a moderate denoising and for  $SNR \leq 10dB$  a strong denoising. Since white noise is assumed, an orthogonal wavelet (Daubechies10-Wavelet) was used.

Figure 1 shows the progression of the T60 estimation. The figure shows the input signal (speech signal in room H2505 with frontal condition) as absolute amplitude in the upper half of the plot. The lower half of the plot shows how many of the subframes of a frame were selected for the T60 estimation of the corresponding frame or if the frame was used at all (corresponds to 0%). A minimum of three out of seven subframes have to be selected in order to assume a sound decay structure in a frame - this corresponds to 43% of the

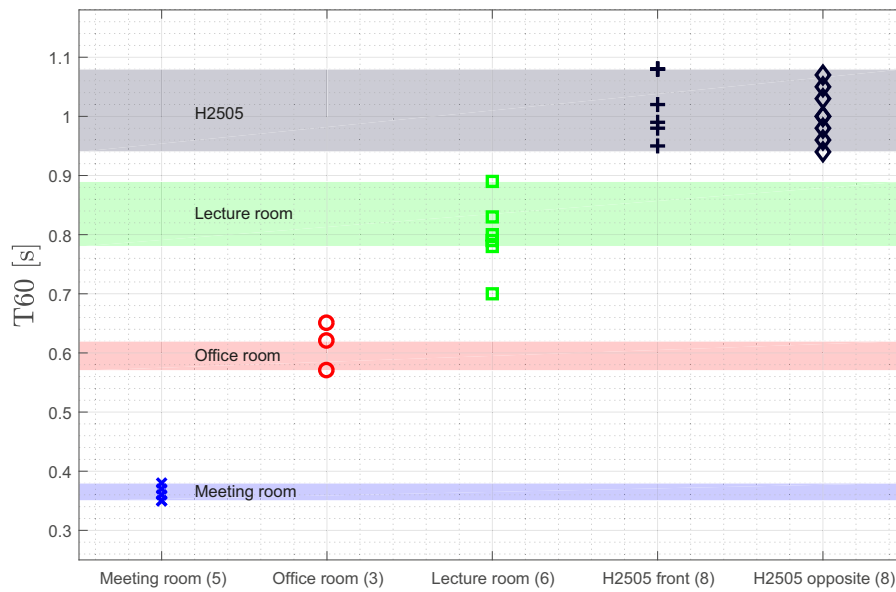
length of a frame. The red line shows the estimated T60 over time. The start value for the T60 value defaults to  $0.5s$ . While the algorithm is real-time capable it takes some time to approach the final T60 estimation.

## 5. Experimental Results

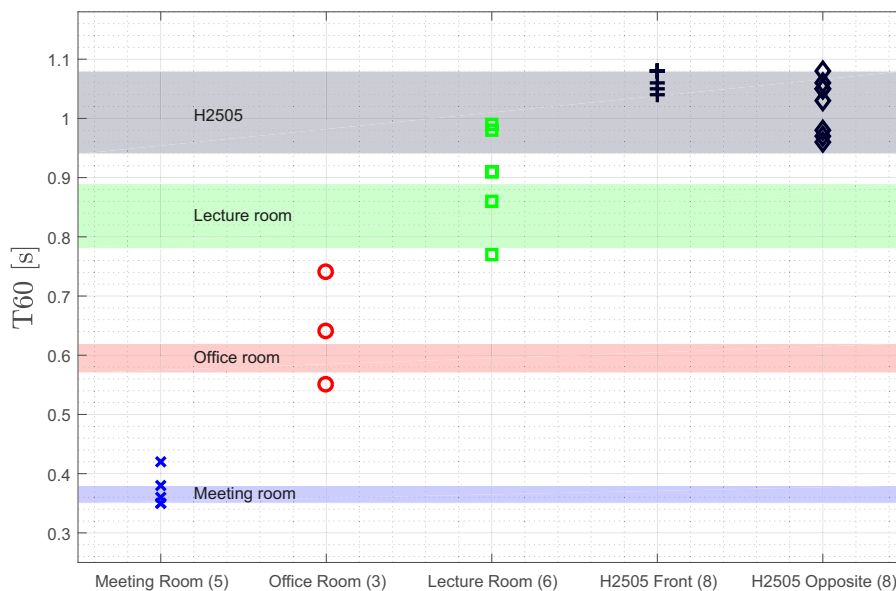
Figures 2, 3 and 4 show the estimation results for the speech signals with different amounts of noise added. In each of the figures estimation results are separated by the rooms. Each marker corresponds to a estimation using one position in the respective room. The amount of available positions is noted in brackets and varies with rooms. The colored areas indicate the range of the T60 values derived from the BRIRs directly, thus indicating the reference range of the reverberation time. Figure 2 shows the condition without additional noise. For the meeting room and both conditions of the H2505 room, all estimations are inside or close to the reference range. For the other two rooms we see several outliers. However, in a classification task those rooms would be easily distinguishable.

However with raising noise levels the estimation gets worse. Figure 3 shows the condition with  $30dB$  SNR and without denoising applied. The estimations for the H2505 room are still within the reference range. For the other rooms the estimations become more scattered.

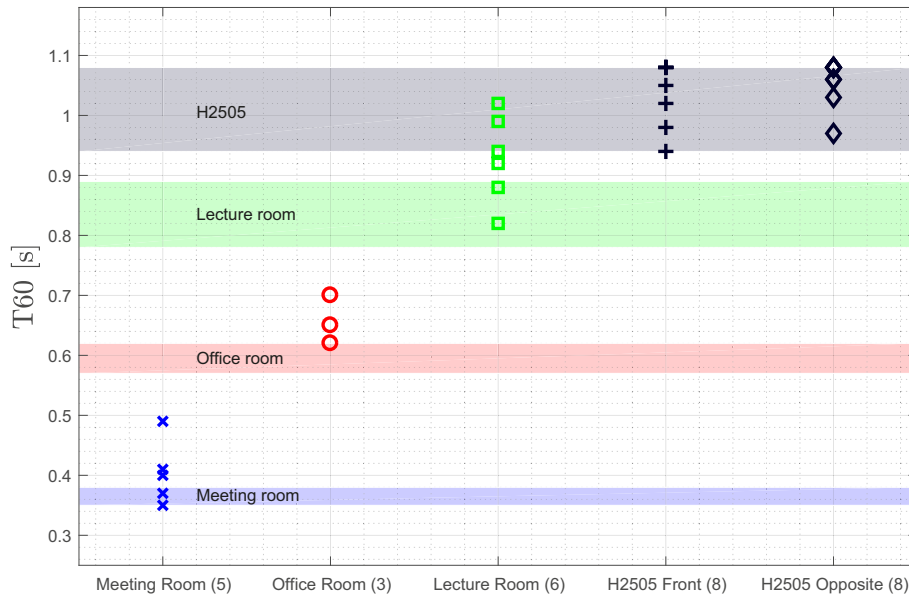
Figure 4 shows the condition with  $20dB$  SNR and with denoising applied. There is a strong tendency towards longer reverberation times for the lecture, office and meeting room. Surprisingly, the results for the H2505 are still very good. The results for the condition with  $10dB$  SNR and with denoising are not shown here, because they are very similar to the  $20dB$  SNR condition and shows the same tendencies. The spread of the estimations is increased again, especially for the smaller



**Fig. 2:** T60 Estimation performance without additional noise. Markers show estimation results for each position in the corresponding rooms. The colored areas indicate maximum and minimum reverberation times in each room derived directly from the BRIRs.



**Fig. 3:** T60 Estimation performance with 30dB SNR. Markers show estimation results for each position in the corresponding rooms. The colored areas indicate maximum and minimum reverberation times in each room derived directly from the BRIRs.



**Fig. 4:** T60 Estimation performance with 20dB SNR and denoising. Markers show estimation results for each position in the corresponding rooms. The colored areas indicate maximum and minimum reverberation times in each room derived directly from the BRIRs.

meeting and office room. Raising noise levels modify the energy decay curve and thus lead to an overestimation of the reverberation time.

### 5.1. Analysis of the DRR

The direct to reverberant ratio is a position and room depended value and therefore a correlation with the T60 estimation was suspected. The DRR is calculated on the basis of the BRIRs and corresponds to the average from left and right. The DRR values in the meeting room range between 8.3dB and 6.8dB, which means a high proportion of direct sound at all positions. In the office room the position depended variation is significant higher. The DRR decreases rapidly with increasing distance from 8.2dB to -1dB. This is due to the many differently reflective surfaces of the room furnishing. This results in a sharp increase in diffuse sound, even with small changes in distance of one meter each. As the diffuse sound component increases, the results of the estimation improves. However, this is only the case for test sounds without additional noise and it must be considered that there were only three measurement positions in this room. In the lecture room the DRR drops sharply from 7.2dB to -5.7dB as the distance from speaker to microphone increases. Accordingly, the direct sound component at the outermost position is considerably lower than at closest position. However, no dependency between DRR and T60 estimation performance can be observed. The T60 estimations in room H2505 were the best among the rooms in this study and it is also the room with lowest DRR values. The DRR in the H2505 room front condition changes from 6dB to -8.5dB with increasing distance. Not surprisingly, the DRR values for the H2505 opposite direction range between -0.5dB and -5.4dB. The DRR analysis could not show a clear dependency between DRR and T60 estimation performance. In some rooms

and noise conditions such a tendency could be observed but other measurements disprove this hypothesis. The estimation performance in room H2505 in connection with the low DRR values leaves room for speculations whether or not there may be an interconnection. A study with more rooms would have to prove this.

## 6. Discussion and Conclusion

In this paper a state of the art approach for blind reverberation time estimation was evaluated for binaural test signals. The estimation for the selected rooms is good to distinguish the rooms. However, this task is not particular difficult given the fairly large differences between the rooms. For test signals without artificial noise the results are close or within the T60 values obtained from the BRIRs directly. With more noise the estimation gets worse, even when denoising is applied. It remains unclear why the estimation performance in room H2505 was the best and the most robust. The DRR could be an explanation, but this could not be confirmed with certainty. The recording equipment and procedure was different between room H2505 and the others, but the authors are in doubt whether this is a decisive factor. The tuning of the parameters of the algorithm require some a priori knowledge about the rooms in order to perform best. Here parameters were chosen which should result in better estimations for rooms with a reverberation time over 0.5s. Another restriction of the algorithm is the delay of the estimations, because the algorithm needs several seconds to approach a stable T60 estimation. It will depend on the application if that is an issue. The algorithm is tuned to speech, because a speech based decay model is used. By combining sound recognition with this approach, different decay models might be applied based on the type of the sound source. Furthermore in future researches additional

acoustical parameters can be used for a better selection of suitable BRIRs. The current state of research can not explain which acoustical parameter is most important with respect to the room divergence effect. A study by Werner [13] found no significant effect of DRR manipulation on externalization in situations of room divergence. This finding draws attention to other features like the temporal structure of BRIRs. If temporal structures are indeed relevant, a possible solution would be to combine an BRIR selection based on T60 with an optical or acoustical geometry estimation. This way BRIRs would be selected from a room which is similar to the actual listening room in terms of T60 and geometry.

Overall, the performance in light of the low computational complexity make this, and other similar algorithms, attractive for further investigations in the scope of augmented auditory realities.

## 7. References

- [1] H. Löllmann and P. Vary, "Estimation of the reverberation time in noisy environments", in Proc. of International Workshop on Acoustic Echo and Noise Control (IWAENC), Seattle, Washington, USA, 2008
- [2] H. Löllmann, E. Yilmaz, M. Jeub and P. Vary, "An improved algorithm for blind reverberation time estimation", in Proc. of International Workshop on Acoustic Echo and Noise Control (IWAENC), Tel Aviv, Israel, 2010
- [3] M. Jeub, M. Schafer and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms", in Proc. of International Conference on Digital Signal Processing (DSP), Santorini, Greece, IEEE, July 2009, pp. 1-4, <https://doi.org/10.1109/ICDSP.2009.5201259>
- [4] F. Lim, M. Thomas, and I. Tashev. "Blur kernel estimation approach to blind reverberation time estimation", IEEE International Conference on Acoustics, Speech and Signal Processing, 2015, <https://doi.org/10.1109/ICASSP.2015.7177928>
- [5] J. Eaton, N. Gaubitch, A. Moore, and P. Naylor. "Estimation of room acoustic parameters: The ace challenge", IEEE Transactions on Audio, Speech, and Language Processing, 2016, <https://doi.org/10.1109/TASLP.2016.2577502>
- [6] N. Gaubitch, H. Loellmann, and M. Jeub. "Performance comparison of algorithms for blind reverberation time estimation from speech", IEEE International Workshop on Acoustic Signal Enhancement, 2012.
- [7] M. Prego, A. Lima, and S. Netto. "Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decomposition", IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2015, <https://doi.org/10.1109/WASPAA.2015.7336954>
- [8] H. Löllmann and P. Vary. "Single-channel maximum - likelihood t60 estimation exploiting subband information". ACE Challenge Workshop, satellite of IEEE-WASPAA, New Paltz, NY, USA, 2015, <https://arxiv.org/abs/1511.04063>
- [9] S. Werner, F. Klein, T. Mayenfels und K. Brandenburg. "A Summary on Acoustic Room Divergence and its Effect on Externalization of Auditory Events". In Proc. of 8th International Conference on Quality of Multimedia Experience (QoMEX). <https://doi.org/10.1109/QoMEX.2016.7498973>. 2016
- [10] K. Brandenburg, E. Cano, F. Klein, T. Köllmer, H. Lukashevich, A. Neidhardt, U. Sloma, and S. Werner, "Plausible Augmentation of Auditory Scenes Using Dynamic Binaural Synthesis for Personalized Auditory Realities", in Proc. of: Conference of the Audio Engineering Society (AES) Audio for Virtual and Augmented Reality, USA, 2018.
- [11] A. Zimmermann and A. Lorenz. 2008. "LISTEN: a user-adaptive audio-augmented museum guide". User Modeling and User-Adapted Interaction 18, 5 (2008), 389-416, <https://doi.org/10.1007/s11257-008-9049-x>
- [12] G. Plenge, "Über das Problem der Im-Kopf-Lokalisation [On the Problem of In Head Localization]", In: Acustica 26.5 (1972), S. 241-252
- [13] Werner, S., Klein, F., and Sporer T., "Adjustment of the Direct-to-Reverberant-Energy-Ratio to Reach Externalization within a Binaural Synthesis System", AES Conference on Audio for Virtual and Augmented Reality, Los Angeles, CA, USA, 2016.
- [14] Z. Ren, K. Qian, Z. Zhang, V. Pandit, A. Baird, and B. Schuller, "Deep Scalogram Representations for Acoustic Scene Classification," IEEE/CAA Journal of Automatica Sinica, vol. 5, no. 3, pp. 662-669, 2018, <https://doi.org/10.1109/JAS.2018.7511066>
- [15] M. Markovic, S. K. Olesen and D. Hammershøi, "Three-dimensional point-cloud room model for room acoustics simulations", in Proc. of Meetings on Acoustics, 19/1, 2013, <https://doi.org/10.1121/1.4800237>





# Sweet area size for the envelopment of a recursive and a non-recursive diffuseness rendering approach

M. Blochberger<sup>1</sup>, F. Zotter<sup>2</sup>, M. Frank<sup>2</sup>

<sup>1</sup> *Graz University of Technology, Austria, Email: matthias.blochberger@student.tugraz.at*

<sup>2</sup> *Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Austria, Email: {zotter, frank}@iem.at*

## Abstract

We compare the extent of the usable audience area of two algorithms that produce diffusely enveloping, multi-channel surround playback from a single-channel input. The FIR approach designs a set of random group-delay allpass filters to generate a set of minimally correlated playback signals. Canfield-Dafilou presented a frequency-dependent maximum group delay value as a constraint to keep audible artifacts small, in studio environments. To enlarge the audience area in which an enveloping and diffuse listening experience is achieved, we relax this constraint while having to accept an unavoidable impression of spaciousness and reverberation. Consequently, the FIR approach naturally competes with IIR feedback-delay network as alternative approach. We conduct listening experiments to reveal quality and effectiveness of both methods, in particular regarding sweet area size and sound quality.

## 1. Introduction

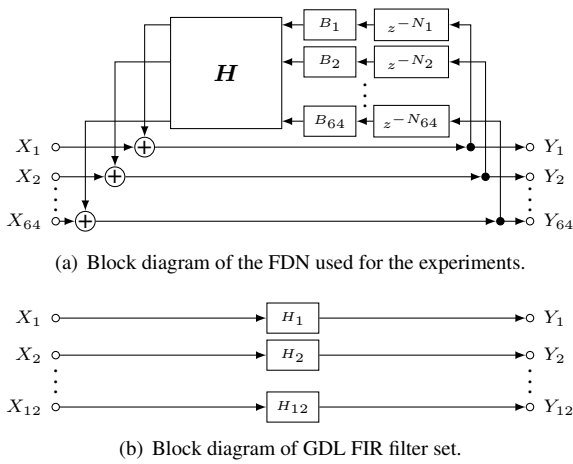
Literature typically defines (e.g. Rumsey [1]) listener envelopment (LEV) as the auditory perception that spacious sound appears to arriving from all the surrounding directions in the space. Often envelopment is defined in contrast to apparent source width (ASW), which refers to the impression of a localized sound, however appearing to be wide. Others refer to measures related to late reverberation to calculate a value for the envelopment as seen in [2, 3].

Multichannel audio playback could use room models, virtual microphones or measurements [4] and auralize them in order to produce the sensation of envelopment. By contrast, this paper regards decorrelation algorithms as less physical concept rendering diffuse sound fields on loudspeakers, such as those presented in [5–10]. Typically, such algorithms produce a set of signals by filtering a single-channel input, so that the set of signals is audio-technically considered to be uncorrelated. The expectation is that feeding those signals to surrounding loudspeakers produces envelopment.

Nevertheless, such algorithms may only partly decorrelate the signals, as audio quality requires to maintain as much of the temporal structure of the original signal as possible. And yet, spacious impressions are often generated as a noticeable side effect. Two candidate algorithms compared here: (i) Canfield-Dafilou recently proposed a promising block-filter implementation designed by random group-delay all-pass filters with frequency-dependent limits [9]. On the other hand, (ii) the feedback-delay network [11,12], is a multi-channel IIR structure offering high onset fidelity and efficiency.

In this paper we are going to illuminate how well the exemplary algorithms perform in producing envelopment over an extended listening area. Zotter/Frank [13] and Frank [14, 15] contained ideas to investigate the sweet-area size of decorrelated/diffuse and enveloping sounds, which we are going to extend, here.





**Fig. 1:** The FDN used in the experiment with the mixing matrix  $H$  in the feedback loop and the GDL FIR approach has 12 filter implemented as FFT.

## 2. FDN (IIR)

Feedback-delay networks are multi-channel recursions, e.g. 64 channels, denoted in the  $z$ -domain as

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{64} \end{bmatrix} = \mathbf{H} \operatorname{diag} \left\{ \begin{bmatrix} B_1 \\ \vdots \\ B_{64} \end{bmatrix} \right\} \operatorname{diag} \left\{ \begin{bmatrix} z^{-N_1} \\ \vdots \\ z^{-N_{64}} \end{bmatrix} \right\} \begin{bmatrix} Y_1 \\ \vdots \\ Y_{64} \end{bmatrix} + \begin{bmatrix} X_1 \\ \vdots \\ X_{64} \end{bmatrix}$$

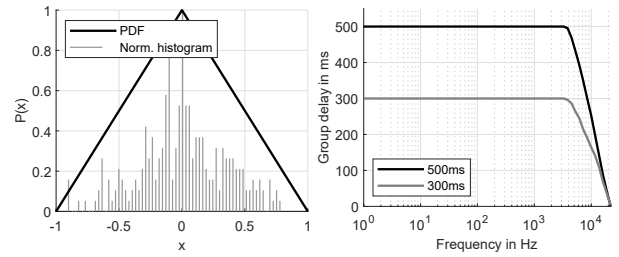
w.  $B_i = g_{lo}^{N_i} B_{lo} + g_{mid}^{N_i} B_{mid} + g_{hi}^{N_i} B_{hi}$ , (1)

their feedback matrix  $H$  is unitary, the channel delays  $N_i$  are individual, and the channel gains in 3 bands  $g_{lo}$ ,  $g_{mid}$ ,  $g_{hi}$  correspond to the desired attenuation per sample (Fig. 1(a)) The network employed in the study is the IEM FdnReverb [16]. It uses a selection of increasing prime numbers to specify the sample delays  $N_i$ . For efficiency and optimal mixing, the unitary matrix  $H$  is implemented as Fast Walsh-Hadamard transform with 64 multiplications (normalization) and  $8 \cdot 64$  sums/differences, see Rochesso [12], which leaves the 3-band IIR filters in the 64 bands as the only costly operation.

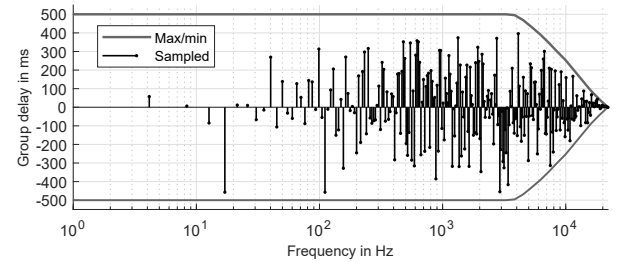
As is, the FDN would start with a strong attack after which the signal decays in a  $10^{-3t/T_{60}}$  shape. In the implementation here, the single-channel input is fed to the inputs  $X_1, \dots, X_{12}$  to get a delayed feed-forward signal to every loudspeaker from the outputs  $Y_1, \dots, Y_{12}$ .

The IEM FdnReverb plugin moreover allows to set a slow onset, which is accomplished by subtracting two 64-channel FDNs. From the main FDN with the desired decay time, another one is subtracted that exhibits a fast decay, which becomes a  $(1 - 10^{-3t/T_{onset}})$ -shaped onset. In the slow-onset implementation here, the single-channel input is fed to the input  $X_1$  to get a delayed feed-forward signal to every loudspeaker from the outputs  $Y_1, \dots, Y_{12}$ .

These FDN input setups were used in all experiments except number 1 (see Sec. 4).



(a) Probability density function (b) Maximum group delay curves



(c) Example of the maximum group delay with the sampled values at ERB-spaced frequencies ( $M = 512, N = 512$ ), in ms.

**Fig. 2:** The triangular probability density function ensures more values close to zero. The maximum group delay curves have a roll off above 4kHz as it provided better sounding impulse responses. The sampled values on the interval  $[-1, 1]$  are scaled to the min/max interval of group delay values.

## 3. Random GDL (FIR)

The FIR approach uses impulse responses generated from random group delay curves in the same manner as Canfield-Dafilou and Abel proposed in [9].

For each filter,  $M$  random values are drawn from a symmetric triangular probability density function (Fig. 2(a)) provided by MATLABs *makedist* method. Having the maximum probability at the zero value ensures a higher amount of zero/small values. These values, spaced on a Moore-Glasberg ERB [17] warped frequency scale, yield the group delay curve. The generated curve is scaled by maximum and minimum values predefined for each frequency (Fig. 2(c)).

To avoid instability of the listening impression stemming from certain phase differences, the group delay curves are modified to limit the differences at low frequencies. One of the generated curves is taken as common ground while the others are scaled to fall into the interval  $[-(4f)^{-1}, (4f)^{-1}]$  in relation to that common curve. A cross fade between the modified curves and originally generated curves happens from 600 Hz to 1400 Hz.

Using any group delay function  $\tau(f)$  over frequency  $f$  the phase is calculated by

$$\phi(f) = -2\pi \int_0^f \tau(f) df. \quad (2)$$

Evaluating the integral for frequencies on the interval  $[0, \frac{f_s}{2}]$  where  $f_s$  is the sampling frequency. Since the positions of the sample values on the frequency scale are not equidistant, a

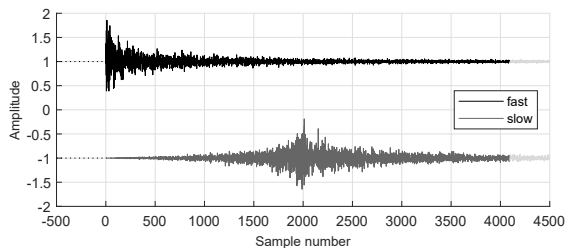
resampling to the equidistant bins  $k = 0, \dots, N/2$  at  $f = \frac{k f_s}{2N}$  has to be done first, in this case linear interpolation provides sufficient precision. Integration is numerically solved by  $\phi[k] = -\sum_{k'=0}^k \tau[k'] \frac{2\pi f_s}{N}$ , and mirroring about the origin yields the phase of the desired skew-symmetric spectrum. Lastly the the inverse fast Fourier transformation yields a real valued impulse response

$$h[n] = \mathcal{IFFT} \left\{ e^{j\phi[k]} \right\}. \quad (3)$$

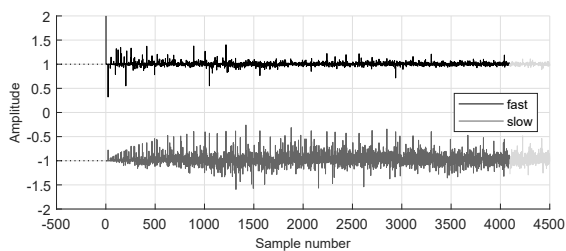
The resulting all-pass impulse responses are symmetric in time which means the slow onsets can be disturbing. An option to mitigate this problem is to truncate the impulse responses at chosen points. The best option for the preservation of transients is to truncate at the center symmetry of the impulse response. For slower onsets it can be truncated at earlier points. In general this leads to better onsets but compromises the all-pass frequency response. The randomized nature of the impulse responses leads to random deviations in the frequency responses, which are cancelling out when using a sufficient amount of impulse responses.

The impulse responses generated by this method and used in this study are of two lengths and onset types. A maximum group delay value curve of 300 ms as well as 500 ms both with a decrease in high frequencies above 4 kHz (Fig. 2(b)) with a fast onset and an onset of 2000 samples (Fig. 3(a)). The slow onset is formed by truncating the symmetric impulse response 2000 samples (45 ms) earlier and multiplying a fade in function of the form  $f(n) = 2\frac{n}{L} - \left(\frac{n}{L}\right)^2$  with the first 2000 samples where  $n = 0 \dots L$  is the sample number and  $L = 2000$  the length of the fade in.

All GDL impulse responses used for the following experiments were generated using the parameters  $M = 131072$ ,  $N = 131072$ ,  $f_s = 44100Hz$ .



(a) Fast and slow onset for a GDL FIR impulse response



(b) Fast and slow onset for a FDN IIR impulse response

**Fig. 3:** Fast and slow onsets for truncated GDL impulse responses and a FDN. The slow onsets for the FDN approach are fit to the peak location of GDL at 45 ms.

## 4. Off-center envelopment (Exp. 1)

Beranek [3] defines envelopment as perceived presence of all sound arrival directions in a room. Choisel/Wickelmaier [18] define it as the perception of a sound that *wraps around you* giving you the impression of *being immersed in it* in contrast to *being outside of it*. Most literature defines it in a similar manner, although a common accepted definition is non-existent.

In our first experiment, listeners were asked which loudspeaker directions did not appear to contribute to the surround sound playback, at two off-center listening positions.

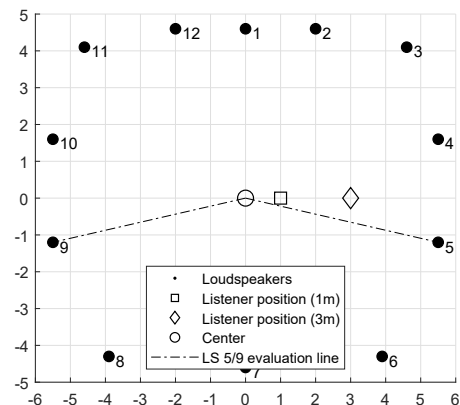
### 4.1. Method

Both FDN and GDL algorithms were used to produce 12 independent impulse responses to feed the 12 horizontal loudspeakers of the IEM CUBE (10×11 m, 500 ms reverberation time; Fig. 4). The parameters of the GDL approach are as described in Sec. 3, the parameters of the FDN are listed in Tab. 1 which were chosen to fit the corresponding GDL conditions by ear. The onset was varied between  $T_{onset} = \{0, 45\}$  ms, yielding 8 conditions in total.

For this experiment only, the inputs of the FDN were fed by a mono source encoded in 5th-order ambisonics in order to feed signal into multiple inputs simultaneously (Azimuth: 24°, Elevation 38°). As a loop sound, *Joyride* from IEM OpenData Archive [19] was used.

8 participants aged from 23 to 39 years took part in the experiment. Every participant did the task alone and had a remote control to switch between the 8 conditions. For each condition, the participant could look into any direction and was asked to write on a piece of paper which of the 12 loudspeaker directions did not appear to contribute at the laterally off-center listening positions -3 m, -1 m, 1 m, and 3 m. 4 listeners were able to finish the task at both, the left and right off-center listening positions within 30 min, and 4 did so for only the left off-center listening position (average time for either left or right positions 22 min).

Listeners reported difficulty in estimating the presence of a loudspeaker direction whenever the perceived sound appeared



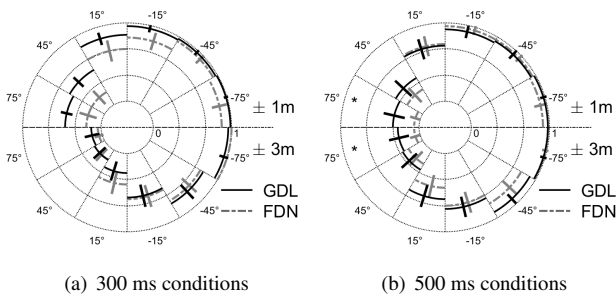
**Fig. 4:** Layout of the IEM CUBE with loudspeaker positions, listening positions for experiment 1 and evaluation lines for experiment 2.

Stim.	Parameter			
	Room size	Rev. Time [s]	Fade-In [s]	Gain [dB]
300ms <i>so.</i>	4	0.7	0.11	0.0
300ms <i>fo.</i>	4	0.7	0	0.0
500ms <i>so.</i>	4	1.0	0.11	0.0
500ms <i>fo.</i>	4	1.0	0	0.0

**Tab. 1:** Settings for the IEM FdnReverb for experiment Nr. 1. The Filter Gain parameter is applied to both bands for a flat filter response. (*so.* = slow onset; *fo.* = fast onset)

Stim.	Parameter			
	Room size	Rev. Time [s]	Fade-In [s]	Gain [dB]
300ms <i>so.</i>	8	0.6	0.10	0.0dB
300ms <i>fo.</i>	8	0.6	0	0.0dB
500ms <i>so.</i>	8	1.0	0.10	-2.6dB
500ms <i>fo.</i>	8	1.0	0	-2.6dB

**Tab. 2:** Settings for the IEM FdnReverb for experiment 2 and 3. The Filter Gain parameter is applied to both bands for a flat filter response. (*so.* = slow onset; *fo.* = fast onset)



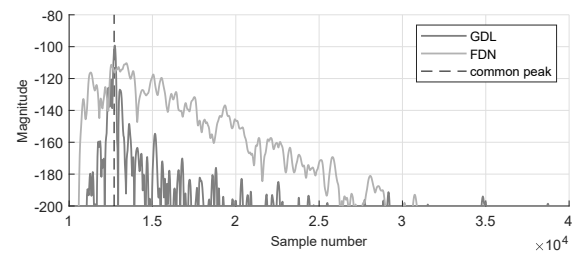
**Fig. 5:** The average perceived loudspeaker activity from binary responses of 12 directions are shown in terms of the means (circular segments) and 95% confidence intervals (radial segments). Sectors with significant differences are marked (\*). Listeners were laterally off-center by either 1 m or 3 m, left and right offsets were mirrored and averaged, and front-back responses were assumed symmetrically pooled (upper half plane: graph for 1 m, lower half plane: graph for 3 m), with short (a) and long (b) decorrelation networks, FDN and GDL.

to be closer than the loudspeaker in distance.

## 4.2. Results

The evaluation was done based on the small set of responses, which was not extended because of the difficulty and duration of the task. As listeners freely changed their look direction, responses for back loudspeakers were mapped to the front, and responses for left-off-center listening positions were mapped to such for right-off-center positions. In this way, 8 responses acquired 12 directions for left offsets and 4 for right offsets, give us  $2 \times (8 + 4) = 24$  responses for 6 frontal directions, per condition. Pair-wise tests on the pooled data indicated hard/soft onset not to be significant as factor, so responses for the hard/soft onset conditions were pooled, yielding 48 data points per direction. By contrast FDN/GDL and short/long response lengths were found to improve the ratings.

Fig. 5 shows the averaged binary directional response (loudspeaker direction perceived to contribute or not) of the 6 frontal response directions on a linear radial scale from 0 to 1. Upper half planes show perceived directional activity for 1 m off-center positions and lower half planes for 3 m. Whereas more reverberant responses of the decorrelator networks generally produce a slight increase in directional activity, we can



**Fig. 6:** The squared impulse responses are summed up to fit the onset peak of FDN to the onset peak of GDL.

also observe in Fig. 5 that distant loudspeakers may only produce a perceived activity up to  $\frac{1}{2}$  according to our results.

For a directional activity rated with  $\geq \frac{1}{2}$ , we can find a coverage of  $\pm 120^\circ$  measured from the closest loudspeaker at 1 m off-center, or  $\pm 90^\circ$  at 3 m. Apparently, already slight off-center positions can exclude distant loudspeakers from a clear audibility in an all-enveloping playback setting.

In greater detail, the GDL algorithm appears produce significantly better results in some cases (Student's T-test). In particular, GDL produces significantly more loudspeaker activity with long group delay time (500 ms) at the  $75^\circ$  direction at both the 1 m ( $p = 0.0001$ ) and the 3 m ( $p = 0.001$ ) off-center position and weakly outperforms at 1 m/ $45^\circ$  ( $p = 0.0882$ ), 3 m/ $15^\circ$  ( $p = 0.0512$ ) and 3 m/ $-45^\circ$  ( $p = 0.0324$ ). The short group delay condition (300 ms) weakly outperforms the FDN counterpart at 1 m/ $-75^\circ$  ( $p = 0.0579$ ), 1 m/ $-15^\circ$  ( $p = 0.0569$ ). On the other hand, the FDN algorithm has a weak advantage at 3 m/ $15^\circ$  ( $p = 0.0702$ ). All other condition-direction combinations show no significant differences.

From these results, we might conclude that envelopment in the standard definition (presence of all directions) is infeasible for an extended audience area in surround playback using individual loudspeakers. While this might appear counter-intuitive, e.g. when regarding the apparent success in [15], plausibility of an enveloping auditory scene might not depend on a detailed presence of all directions. The presence of dry and direct as well as lateral reverberant sound might already be satisfactory.

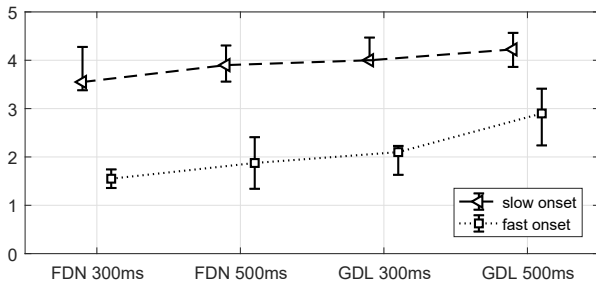


Fig. 7: Rating of the lateral limit of the plausible-reproduction sweet area in meters, with regard to direct frontal and reverberant enveloping/lateral sound, in medians and 95% confidence intervals.

## 5. Sweet-area size (Exp. 2)

As envelopment might be too strict a criterion to evaluate the sweet-area size, our second experiment considers direct sound from the frontal loudspeaker together with the diffuse sound from all the 12 horizontal loudspeakers.

### 5.1. Method

The 8 conditions were the same GDL impulse responses as in experiment 1 as described in Sec. 3 and the FDN parameters as listed in Tab. 2. The direct sound from the frontal loudspeaker was set at a level of  $-6$  dB compared to the diffuse sound field at the center position.

For this experiment, the parameters of the FDN implementation were fit to the GDL approach by the measure of the center time

$$T_s = \frac{\int_0^\infty t \cdot h^2(t) dt}{\int_0^\infty h^2(t) dt} \quad (4)$$

for the fast-onset stimuli and additionally fitting the onset peak of the FDN to the onset peak of GDL at 45 ms (Fig. 6). Again, the onset was varied between  $T_{\text{onset}} = \{0, 45\}$  ms, yielding 8 conditions in total and as a loop sound, *Joyride* from the IEM OpenData Archive [19] was used.

Listeners were individually asked to switch through 8 conditions with a remote control. For each condition, their task was to write on paper the lateral limits of the sweet area when walking the line from the central listening spot towards either loudspeaker 5 (at  $110^\circ$  right) or 9 (at  $110^\circ$  left) as depicted in Fig. 4. We gave them a definition of the sweet area as being laterally limited where either (i) the frontal sound would move outside the  $\pm 30^\circ$  range of the frontal loudspeakers 12-1-2, or (ii) the lateral reverberation begins to dominate and disturb the spatial impression.

8 persons between 24 and 55 years of age participated in the experiment with an average duration of 12 minutes. The lower time it took participants to complete this task, suggests that this task was much simpler than the previous one.

### 5.2. Results

Fig. 7 shows that overall that diffuse impulse responses with slow onset are most effective when desiring an extended sweet area for direct sound plus reverberation. A Wilcoxon

signed rank test with Bonferroni-Holm correction confirms the significantly larger sweet area for each condition ( $p < 0.009$ ). Only in case of the GDL with fast onset, long reverberation contributes to a enlarged sweet area ( $p = 0.025$ ). In detail, the short GDL with slow onset weakly outperforms its FND counterpart ( $p = 0.051$ ). The same holds for the long GDL with fast onset ( $p = 0.087$ ).

## 6. Envelopment/transients (Exp. 3)

While the experiments above did not consider audio quality aspects and would not give much insight yet into which of the algorithms is more effective, our third experiment uses a multi-stimulus test setup for the central listening position (Fig. 4) to acquire a closer differentiation.

### 6.1. Method

The conditions were chosen as in the previous experiment investigating sweet area size, however without the frontal direct sound. The 12 surrounding loudspeakers were fed by the 8 different filter sets to switch between and the listeners were asked to comparatively rate in multi-stimulus trials: (i) the preservation of transients (min...max) as an audio quality aspect, and (ii) the diffuse envelopment (min...max) they perceived as a criterion of effectiveness. Both multi-stimulus tasks were rated twice per listener, each time with randomly arranged assignment of the stimuli to the 8 sliders. Like previously, *Joyride* from IEM OpenData Archive [19] was used.

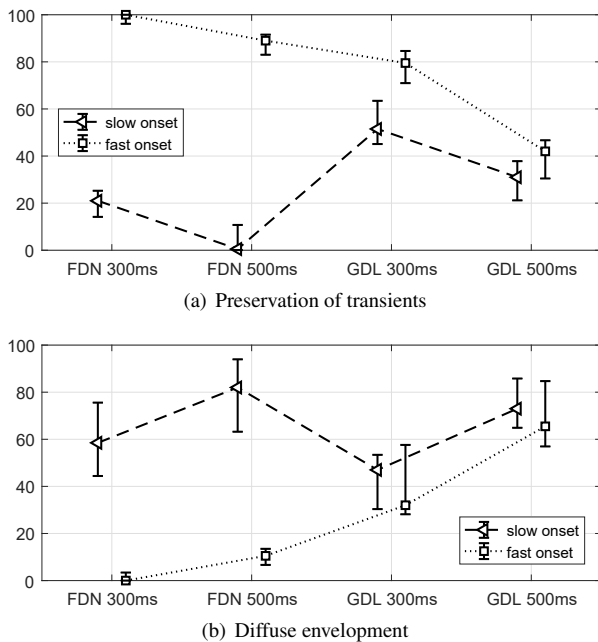
As this experiment happened in conjunction with the previous one, the same 8 persons aged 24 to 55 years participated. The average completion time was 18 minutes.

The participants reported difficulty in the diffuse envelopment task, especially whenever conditions appeared to be different in timbre (high frequencies) or onset, which occasionally made their decision difficult.

### 6.2. Results

Fig. 8(a) shows that the slow onset deteriorates the perceived preservation of transients for both FDN and GDL ( $p < 0.01$ ), except for the long GDL ( $p = 0.125$ ). This negative effect is worse in the FDN conditions compared to such with GDL. In general, quality decreases for longer decay times ( $p < 0.011$ ). The fast onset conditions (dotted) are generally of higher quality, and there, the FDN conditions outperform the transient preservation of GDL ( $p < 0.011$ ), especially for the long decay time. This might be due to the linear decay profile of the GDL responses which on the other hand appears to be beneficial under slow onset conditions, as GDL largely outperforms FDN in terms of diffuseness.

The comparison of the diffuse envelopment rating of the sound fields in Fig. 8(b) shows an increase with the decay length ( $p < 0.029$ ) and increase with onset time by tendency: In case of the FDN approach, the fast onset conditions perform significantly less effective ( $p < 0.011$ ) than their slow counterparts. However, in case of the GDL approach, there



**Fig. 8:** Comparative rating of the diffusion networks in terms of quality (preservation of transients) and effectiveness (diffuse envelopment) at central listening position concerning medians and 95% confidence intervals.

are no significant differences for both the short and long decay time.

The GDL condition with 300 ms with slow onset appears to be a good compromise between preservation of transients (quality) and the perceived diffuse envelopment (effectiveness), and despite the larger efforts required, the GDL approach appears to be an interesting alternative.

## 7. Conclusion

In this contribution we compared two fundamentally different diffuseness and envelopment rendering approaches on a rather large loudspeaker setup, random-group-delay (GDL) allpasses implemented as FIR structures and feedback-delay networks (FDN) that are implemented IIR.

Independent of the algorithms and their settings, we found that it is challenging to produce diffuse envelopment across a large audience area in experiment 1 (Sec. 4), in particular in the strict understanding of sound being perceived to directionally arrive from everywhere. This is because the contribution of distant loudspeakers soon becomes imperceptible, even at relatively small distances to the central listening position. The standard definition of envelopment appears impractical in evaluation of sound fields for larger audiences. As a refinement of the method, future experiment might consider asking for the auditory distance of the enveloping sound into a given set of directions.

For diffuse rendering with direct sound, experiment 2 (Sec. 5) showed that independent of the approach, methods with slow onset have a crucial advantage over such with fast onset. Consistent rendering of envelopment is possible for a sweet

area of about 4 m when using algorithms producing impulse responses with mentioned slow onsets, in contrast to such with fast onsets (2 m). Moreover, the random group-delay-based approach appears to produce a slightly larger sweet area.

For both algorithms, we showed in experiment 3 (Sec. 6) that they produce more envelopment when used with slow onset, in particular the FDN structure. Additionally, the proposed GDL structure with moderate time constants (300ms) can be recommended in terms of its pronounced effect, while its transient preservation is quite acceptable.

## Acknowledgements

We thank the voluntary listeners for their participation in our experiments.

## References

- [1] F. Rumsey, "Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm," *J. Audio Eng. Soc.*, vol. 50, no. 9, pp. 651–666, 2002. [Online]. Available: <http://www.aes.org/e-lib/browse.cfm?elib=11067>
- [2] G. A. Soulodre, M. C. Lavoie, and S. G. Norcross, "Investigation of listener envelopment in multichannel surround systems," in *Audio Engineering Society Convention 113*, Oct 2002. [Online]. Available: <http://www.aes.org/e-lib/browse.cfm?elib=11283>
- [3] L. L. Beranek, "Concert hall acoustics," *J. Audio Eng. Soc.*, vol. 56, no. 7/8, pp. 532–544, 2008. [Online]. Available: <http://www.aes.org/e-lib/browse.cfm?elib=14398>
- [4] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, "Spatial decomposition method for room impulse responses," *J. Audio Eng. Soc.*, vol. 61, no. 1/2, pp. 17–28, 2013. [Online]. Available: <http://www.aes.org/e-lib/browse.cfm?elib=16664>
- [5] E. Weinstein, M. Feder, and A. V. Oppenheim, "Multi-channel signal separation based on decorrelation," Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, RLE Technical Report 573, 1992.
- [6] G. S. Kendall, "The decorrelation of audio signals and its impact on spatial imagery," *Computer Music Journal*, vol. 19, no. 24, 1995.
- [7] G. Potard and I. Burnett, "Decorrelation techniques for the rendering of apparent sound source width in 3d audio displays," in *Proc. of the 7th Int. Conf. on Digital Audio Effects*, 01 2004, pp. 280–208.
- [8] M. Boueri and C. Kyriakakis, "Audio signal decorrelation based on a critical band approach," in *Audio Engineering Society Convention 117*, 01 2004.
- [9] E. K. Canfield-Dafilou and J. S. Abel, "A group delay-based method for signal decorrelation," in

- Audio Engineering Society Convention 144*, May 2018. [Online]. Available: <http://www.aes.org/e-lib/browse.cfm?elib=19508>
- [10] S. Schlecht, B. Alary, V. Välimäki, and E. Habets, “Optimized velvet-noise decorrelator,” in *21st International Conference on Digital Audio Effects (DAFx-18)*, Aveiro, Portugal, 09 2018. [Online]. Available: [http://dafx2018.web.ua.pt/papers/DAFx2018\\_paper\\_23.pdf](http://dafx2018.web.ua.pt/papers/DAFx2018_paper_23.pdf)
- [11] J. Stautner and M. Puckette, “Designing multi-channel reverberators,” *Computer Music Journal*, vol. 6, no. 1, pp. 52–65, 1982.
- [12] D. Rocchesso and J. O. Smith, “Circulant and elliptic feedback delay networks for artificial reverberation,” *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 1, pp. 51–63, Jan 1997.
- [13] F. Zotter, M. Frank, M. Kronlachner, and J. Choi, “Efficient phantom source widening and diffuseness in ambisonics,” in *EAA Symposium on Auralization and Ambisonics*, Berlin, 2014.
- [14] M. Frank and F. Zotter, “Spatial impression and directional resolution in the reproduction of reverberation,” in *Fortschritte der Akustik - DEGA*, Aachen, 2016.
- [15] —, “Exploring the perceptual sweet area in ambisonics,” in *Audio Engineering Society Convention 142*, Berlin, 2017. [Online]. Available: <http://www.aes.org/e-lib/browse.cfm?elib=18604>
- [16] Plug-in suite by iem. Software Package. Institute of Electronic Music and Acoustics. [Online]. Available: <https://plugins.iem.at/>
- [17] B. Moore and B. Glasberg, “A revision of zwicker’s loudness model,” *Acta Acustica united with Acustica*, vol. 82, pp. 335–345, 03 1996.
- [18] S. Choisel and F. Wickelmaier, “Evaluation of multi-channel reproduced sound: Scaling auditory attributes underlying listener preference,” *The Journal of the Acoustical Society of America*, vol. 121, pp. 388–400, 02 2007.
- [19] Opendata archive. Archive. Institute of Electronic Music and Acoustics. [Online]. Available: <https://opendata.iem.at/>







# Abstract Reviewed Paper at ICSA 2019

Presented \* by VDT.

## Mixed Realities: a live collaborative musical performance

A. Genovese, M. Gospodarek, A. Roginska  
*New York University, USA, Email: genovese@nyu.edu*

### Abstract

In the presented work, a live-rendered percussionist is transformed into a virtual game character and performs a piece along digital avatars created from recordings (audio and motion-capture) of other members of an ensemble, while audience members can observe the collaborative performance through VR headsets. To create a cohesive and compelling result, the auditory expectations of the listeners need to be considered in terms of acoustic integrity between real and virtual sources and spatial impression of each avatar performer.

This paper presents an overview of the workflow and motivations behind this pilot experiment of a novel musical experience, laying down the foundations for future subjective studies into collaborative music performances using virtual and augmented reality headsets. Particular focus is given to the technical challenges concerning the audio material, the perspectives of each participant role, and the qualitative impressions of musicians and audience.

## 1. Introduction

The question of digitally augmented music collaboration is ever-relevant in the field of immersive audio and musical performance. Currently, the adoption of commercial immersive headset devices for virtual and mixed reality can enable today's musicians to experience enhanced forms of virtual presence when connected to their peers. For instance, motion-capture data of a human being can be streamed and live-rendered using complex camera-sensor systems. Game avatars animated by real human beings are thus brought into a shared virtual space where observers and participants are part of a common world. A performer's captured audio stream can be paired with associated metadata to create a three-dimensional trajectory of an audio source, or to create an association to a specific digital element, or avatar, visualizable in the shared virtual scene via a head-mounted display device (HMD, or headset). Usually, the main goal of this application is to create an illusion of realism (or hyper-realism) for the participants involved while maintaining an effective musical output quality. In other words, to achieve a sensation where the collaborating musicians are perceived, by themselves or

by an audience, as if playing together in the same room.

Such prototype systems have been recently explored for immersive theatre applications where live or recorded actors are experienced virtually by a co-located audience [1, 2]. Musical environments have also been explored by testing designs of musical interfaces [3, 4] or assessing the factors that can improve the perceived experience plausibility [5] and co-presence [6, 7]. However, music-making and concert experiences in the traditional sense have so far been more complex challenges, not extensively covered in literature.

As these technologies become progressively more available, it is worth to explore the implications for new kinds of enhanced music performances. Using HMDs, musicians and audience members alike can be placed in a shared virtual world while being, for instance, in physical remote locations. Another avenue is given by Mixed Reality (MR) scenarios, where digital elements are rendered with a degree of connection to the physical world of a user [8]. MR is interesting as it raises a question of considering the individual perspective of each participant in terms of the features that define their local reality both visually and acoustically. In MR applications,

the real and the virtual elements are treated to blend together into a cohesive scene creating an illusion of an augmented world where virtual objects assume location-specific properties, thus seeming more "realistic". In terms of audio, this involves acoustic character of the space, (early reflections, late reverberation, etc.), characteristics of the audio source (e.g. radiation patterns) and spatialized, dynamic, 3D localization of individual sources.

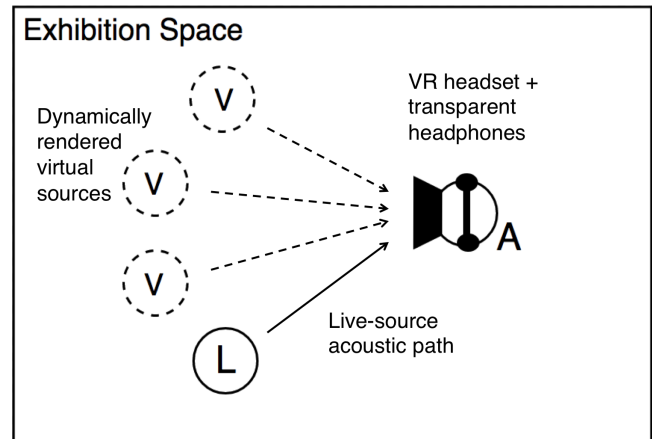
The number of variables and possible combinations involved into organizing immersive mixed-reality performances is only bounded by the amount of technology at the disposal of the artists and engineers. Since it is a complex, if not impossible, task to generalize a qualitative study for all kind of virtual music network topologies, it is usually necessary to limit the investigation by having a target user and scenario in mind.

We hereby present a pilot implementation of a novel mixed-reality musical experience using motion-capture and VR headsets, meant to serve as a first test and canvas upon which to build a series of formal studies on the subject. Although only non-formal assessment was conducted for this work, the establishment of a workflow pipeline helped to investigate an exploratory scenario and provide insights on future evaluation methods. This paper discusses the design, implementation and technical challenges of this work, with particular focus on the audio elements.

In the presented work, a dancer and three percussionists of an African music ensemble have been individually recorded with microphones and an OptiTrack motion capture system. Their audiovisual data was then converted into digital avatars able to be reproduced as virtual performer animations through VR headsets. During an exhibition demo, an additional live percussionist, wearing a motion-capture suit, was brought into the virtual scene as a real-time character. A single audience member, located in the same space as the live-percussionist, was able to observe the joint collaborative performance of all musician avatars, live and pre-recorded, by means of a VR headset and headphones.

## 2. Experience Design

For this particular piece, it was decided to explore a mixed-reality setting where a combination of live and pre-recorded musician game avatars (created with audio and motion capture data) would play together and be observable within a VR headset by an audience member. Within this framework, a number of possibilities were open for exploration in terms of organization and disposition of the three types of participants involved (audience, live musician and pre-recorded musicians, from here referred to as the "virtual ensemble"). What was of interest was the question of how to design a convincing sense of "co-presence" and how to create a sense of shared reality from the perspective of the target user, in this case, the audience. To achieve this, it was decided to tailor a mixed-reality (MR) experience grounded in the local auditory reality of the audience. In other words, fit the virtual material to adapt to a predefined "concert" space shared by the both the live performer and



**Fig. 1:** Design for the spatial disposition of participants during the exhibition phase. The live musician (L) shares the room with the audience (A). The virtual avatars (V) are spatially located around the listener to form a virtual ensemble with the live performer. The audience is provided with transparent headphones in order to allow the local acoustic path to be heard with little obstruction while dynamically rendering in binaural format the sound of the virtual members (acoustically treated).

audience participants. Within this shared space, it was desired to make the experience "feel like a concert" where all the musicians could be perceived as if "being together in the same space" [9]. Using this *audience-centric* approach, all the elements had to be put into service for the enhancement of the audience perspective and quality of experience, but also without compromising the performance of the musician.

Mixed-reality experiences have a different set of challenges in comparison to pure virtual reality experiences. Having a sonic reference from the real world allows the auditory system to compare the real and virtual cues and base the judgment of "scene realism", or plausibility, upon the correlation between the two. In this scenario, as the live musician shares the concert space with the listener, the sound emitted by the "local" instrument and its natural acoustic room reflections, is heard by the audience. This forms the perceived ground reality to which the "remote" virtual sound needs to adapt towards. For the virtual content to mix and adapt cohesively to the concert space and the local acoustic character, artificial reverberation using a room-impulse-response (RIR) measured *in-loco* can be overlaid through convolution to each object source, provided the source-material is recorded in dry condition. This step ensures a shared acoustic character to all elements, helping to build the auditory integration of the scene [10], while small mismatches of that character could instead potentially decrease that sense of integration.

From the perspective of the audience, the live sound needs to be heard naturally as it fits with the visual display, whether real or virtual, thus unobstructed by closed surfaces between the ears and the direct path of the source. In regards to the virtual sources, a dynamic binaural rendering process can create externalized, localizable, virtual sound emitters, which help the listener to build a cognitive organization of the stage display. These considerations can be addressed by the use of

	VE	LM	Aud.
Auditory Env.	Virt.	Real	Real & Virt.
Visual Env.	Virt.	Real	Virt.

**Tab. 1:** Table of shared auditory and visual environments between the three participants: virtual ensemble (VE), live musician (LM), and audience (Aud.). The labels indicate "Real" environment (Re.) or "Virtual" environment (Virt.).

open-back headphones connected to a binaural engine and a head-tracking device. Open-back headphones allows for local real sound sources to be perceived reasonably uncolored, while the integrated IMU units in modern HMDs can provide the positional data for six-degrees-of-freedom movements of a user. In fact, when the 3D binaural rendering is made responsive to rotations and position shifts it dramatically enhances the spatial impression, presence, and perceived directional accuracy [11]. Ideally, a transparent, or hear-through headphone device would be better than the open-back kind as it provides the least amount of coloration of free-field sources [12]. The use of a loudspeaker system is theoretically also possible but it makes it harder to accurately reproduce the directivity patterns of the instruments.

The exhibition of the performance was thus planned as portrayed in Fig. 1. The local presence of a real sound source in proximity to the listener, in conjunction to the delivery of spatialized object-based audio via headphones, effectively creates an auditory mixed-reality scenario on top of which, a virtual visual environment has to be projected on a VR HMD. The resulting dynamic is that of a combination of shared environments between the three types of participants (Tab. 1). Most interestingly, the audience shares the auditory space with the live musician, but the visual space of the virtual ensemble (including a live-rendition of the live musician). To achieve congruence in any realistic display, the bond between the visual and auditory senses needs to be considered; a listener's expectation of the acoustic character of a space is in fact influenced by its visual impact [13]. Cohesiveness between the two sensorial realms is key for achieving a convincing mixed-reality base of display and avoid poor engagement with the intended application. In practice, as a VR headset is here needed to display the avatars and provide the head-tracking data, the virtual visual environment projected to a viewer needs to be congruent with their experience of the local auditory reality if a compelling blend is desired between the live drummer and the virtual ensemble. As we grounded this experience based on the local sound, the visuals need to adapt by displaying an environmental location which would plausibly relate to the local acoustics.

### 3. Implementation

The implementation of this experience can be divided into three main phases, data capturing, scene design, and exhibition. While the capturing phase was conducted to collect material also usable for possible independent projects, the material was selected to be specifically appropriate for the

exhibition described in this paper.

The selection of musical content for the performance had to take into consideration the use of motion capture sensors suits which may interfere in the regular use of an instrument. The selection of an African percussion quartet (Djembe), plus a dancer, was deemed appropriate as the suit proved to be non-invasive to the musicians and also helped to later simplify the graphical rendition of the instrument in the virtual scene. Percussive, highly-transient, content is also more easily localizable in binaural displays as its wide frequency bandwidth covers the range of both ITD and IID cues in generic HRTF sets [14]. The presence of the dancer was effective in raising the aesthetic value of the final piece but was not a determinant element of the exhibition presented.

#### 3.1. Motion and audio capture

The capturing session was performed in a medium size room equipped with optical cameras for motion tracking. Each of the musicians was recorded separately to ensure full control over each take. The goal was to obtain individual, dry recordings of each member of the ensemble so that each one could be treated as a separate audiovisual object.

The motion tracking was performed using an OptiTrack system with 15 cameras and the Motive capture software [15]. Each percussionist, and the dancer, were fitted with a 32-sensors suit, including gloves. As the suit did not seem to impair the musical performance ability, the audio recording could have happened simultaneously to the motion-tracking. For recording the sound of the Djembe drum, four condenser microphones with cardioid pattern were used. This directivity pattern was chosen in order to achieve a good separation of the direct sound to the recording room reflections. The top two microphones were oriented towards the drum membrane, while the bottom two were placed close to the resonance chamber. The ambience sound was also captured using the first order Ambisonics microphone Sennheiser Ambeo, positioned in front of the drum. The ambience sound was only collected for reference and future use, and not included in later parts of this production.

The main drummer was recorded first using a metronome click. The best take was chosen and the track was used as a reference for the following drummers who were recording the other voices of the percussion piece. In total, three drumming parts and a dancer part were recorded, leaving the last drummer part for the live exhibition.

Since it is common, when capturing full body motions, that sporadic glitches may occur for some of the sensors, the motion-tracking data had to be passed through a "cleaning" procedure before further editing. This consisted in correcting the position of each individual sensors at the frames where those glitches resulted in an unnatural skeleton rendering.

#### 3.2. VR Scene Design

The NYU Future Reality Lab provided the necessary facilities that the project required, a large enough room for the performance needs, and a motion capture system for the live-

musician. All the steps described in this section involved a tailored approach to this particular space.

### 3.2.1. Visual Design

An additional important reason for the selection of the room, was the availability of a 3D model rendition of the actual performance space as a digital asset. This made it possible to provide to the audience a semi-identical virtual environment to the actual physical one, thus optimizing the coherence of the local sound with the visual stimulus.

The VR scene was built using the Unity game engine [16] using the aforementioned laboratory asset and digital models of African percussion instruments. The cleaned and synced motion-capture data was used to create skeleton rigs in the AutoDesk Maya software [17] which were used to animate digital characters back in the Unity scene. The three virtual percussionists were disposed as shown in Fig. 1 while the dancer figure was placed in the background behind the digital ensemble.

An Optitrack tracking system was mounted in order to capture the live musician during the exhibition and transform the motion into an additional character in Unity (live-streaming was implemented through the Motive asset package). A calibration step was performed in order to achieve a one-to-one spatial match between the digital rendering of the live musician (plus a digital drum model) and its actual physical position in the room. This was important in order to ensure the perfect match between localization of visual avatar and the live free-field sound of the performer.

### 3.2.2. Audio rendering

To obtain an acoustic characterization of the space, measurements of the room acoustic impulse response were retrieved from an earlier project conducted in the same location. Two omnidirectional microphones DPA 4006 were mounted in the center of the performance space 17 cm apart and at the height of 1.8 m. The measurement signal was reproduced by one speaker located 3 m from the microphones. The ScanIR (v2) MATLAB toolbox [18] was utilized to reproduce the measurement signal and capture the impulse responses. A 2-second sinesweep was thus recorded in stereo at 96 kHz, ranging from 20 Hz to 20 kHz [19]. The impulse responses were later used to process the signal and superimpose the acoustic characteristics of the exhibition space onto the rendered audio tracks, increasing the timbral blend between performers.

For each of the musicians, two microphone positions, one from the bottom and one from the top of Djembe drum were selected out of the four available, and rendered as separate audio tracks. A gentle compression was applied to achieve a more satisfactory timbre of the instrument. The tracks were used to create virtual audio objects implemented in Unity within the prepared visual scene. The Steam-Audio plugin [20] was employed as the audio rendering engine. Each virtual Djembe model was assigned two object sound sources, one for each of the two top-bottom microphone tracks. The transient slap sound from the hand hitting the membrane was thus placed at the top of each digital

instrument while the low frequency resonance was positioned at the bottom. This double-emitter strategy ensured that the size of the instruments and their radiation characteristics were preserved.

The scene spatialization was handled by the binaural rendering engine of Steam Audio, using generic HRTFs and the rotation/position data from the VR headset IMU unit and its tracking sensors, which allow for dynamic perspective. The location of each of the object sound sources was rendered according to its avatar position in relation to the viewer, while distance attenuation was simulated by use of the square-law [21].

In order to apply the diffused reverberation of the room, all the microphone tracks were summed together and the resulted signal was convolved with the late diffused part of the stereo impulse response earlier collected. Finally, the resulting reverberant stereo file was mixed with the dry binaural mix of the avatars (where a slight EQ was applied), in order to achieve a calibrated balance of direct vs reverberant sound deemed aesthetically satisfactory for a compelling experience.

### 3.3. Exhibition

The experience demo was exhibited to a crowd of academics during an internal event. Rehearsals with the live musician were first conducted in order to test the system and allow sufficient comfort in performing while wearing a tracking suit (Fig. 2). It is worth to note that the musician was also part of the original motion-captured ensemble, meaning that there was familiarity with all the parts in the piece and with performing with the suit. The live performer was not provided with a separate VR headset, but with the binaural audio stream deriving from the movements of the audience device. This was not ideal in terms of providing the musician with an optimal auditory perspective but ensured that a perfect synchronization with the recorded material could be maintained. Because a single room impulse response was available, the audio levels had to be mixed for a single seating location of the audience in the space. The audience seat was placed in front of the virtual ensemble and the balance between direct



**Fig. 2:** Video still of the exhibition rehearsal. The VR audience POV of the audience is shown in the background picture, while the overlaid smaller picture illustrates the external view of the live musician and the audience, seen from the experimenter. A video of the exhibition rehearsal is available at <https://www.youtube.com/watch?v=-0VqIn1pTA0>.

and reverberation sound was adjusted for that location. The audience (one person at a time) was fitted with a tethered VR headset and open-back headphones, and encouraged to look around and shift their head position within the area of their seat, but to not walk around the performance space. No formal questionnaire data was collected from the audience, but the musician participants had a chance to share their impressions with the authors.

## 4. Discussion and future work

The work described in this paper was exhibited in an informal setting closer to an art show rather than a formalized experiment. However, a lot of insights were gathered as well as the consolidation of the design criteria that lead to this kind of implementations.

Although the design was audience-centric, no particular audience feedback was received other than general comments about the visual quality of the meshes. However, the live performer was able to respond to a short post-performance questionnaire made of three scale ratings and an open-ended feedback form. The musician rated the comfort of performing while wearing the suit as 3 out of 7 (1 being "Very uncomfortable" and 7 being "Very comfortable"). The sense of acoustic cohesiveness between real and virtual sound was rated 5 out of 7 (1 being "Not cohesive at all" and 7 as "Very cohesive"). Finally, the difficulty of playing with avatars rather than real life musician was rated to be 3 on a seven-point likert scale (1 being "Easier", 7 being "Harder", with 4 being "No Difference"). Other general impressions included the fact that the performance felt like a "one-way avenue of communication, where my job was to fit myself into this world that was created for the experience" and it "did not feel as organic as performing with other people in real time, perhaps because of some visual aspects". Naturally, this is a single data point which needs to be further explored before generalizing to wider settings.

These answers will be taken into consideration for future work, and compared to possible situations where instead all the participants will be live. Some improvements might be achieved in future by treating the live musician point-of-view with the same approach as the audience perspective (acoustic adaptation and binaural dynamic response from their own perspective). Our project also assumed that the audience members were not changing their position drastically since the implemented impulse response was captured at a single point in space with an omnidirectional mic. In future implementations, adding more measurements and dynamic controls for the direct sound to reverberation ratio could add a further degree of realism to the scene and give more flexibility of movement for the audience. More accurate room impulse responses would be obtained by adjusting the source-emitter position at the intended locations of each participant, adding a more accurate early-reflection pattern rather than just the diffused part.

Having built this pilot framework, future studies on mixed-reality and music performance will be conducted in order to investigate the technical and cognitive aspects which regulate

the subjective quality of experience. When talking about evaluating and measuring such quality of experience, it is important to differentiate the metric according to the role of a participant. A setting can be indeed defined in terms of the *target-user*, indicating if the outcome needs to be compelling to the musician for a music-making experience, or to an audience for a concert. The perspectives of the audience and the live-musician could be both taken into account into deciphering the success of musical virtual environments. While they both might tie quality to their sense of presence into the scene, the musician might seek something more keen to an intersection of "co-presence" and "naturalness", as in the sense of "being performing together" to the fellow performer, in a setting comparable to real life. The evaluations from the two perspectives might or might not correlate. However, some kind of different design choices oriented towards *hyper-realism* or *unrealism* might have to be evaluated not in terms of naturalness (or realism), but in terms of telepresence and plausibility, which connect respectively more with a general sense of engagement and coherence between not-necessarily-real environments.

Qualitative observations through questionnaires might reveal how some of these subjective attributes vary according to the nature of the content and the network topology created (e.g. distribution of musicians between local and "remote", or live and virtual). Furthermore, some alternative strategies such as task-success metrics (e.g. correctness of musical output, musical synchronization ability, etc.) or parametric control by the audience [5], could provide a more objective measure and reveal the relationship between technical aspects, perceived quality and effective outcome.

The outlook of future applications related to this work its oriented towards studying virtual and mixed reality for distributed music networks. Music collaboration over the internet has been explored for years [22] but only now its possible to explore virtual performance spaces and connect these endeavours with a three-dimensional virtual presence of the musical performers. In this field, latency is still the primary concern, and streaming motion-capture data efficiently over the internet is a challenge yet to be overcome. However, putting this issue aside, it is still worth to study all the other aspects which relate to these systems.

In our case, the use of pre-recorded material for the guest musician to perform a part, can allow a test system to bypass the issues of signal latency that are inherent in music network collaborations. This also helped to focus the effort on the experience design and reduce rehearsal times for the musicians involved. In future iterations of this experience, it is intended to simulate distributed settings where all musicians are live in order to fully study a real-time collaborative mixed-reality music environment in either musician-centric or audience-centric evaluation schemes.

## 5. Acknowledgments

The authors would like to thank all people who helped in the building of this demo. Christopher Allen O'Leary (drummer and live performer), Max Meyer (drummer and dancer), Jared



Shaw(drummer), Sripathi Sidhar, Christy Welch, Scott Murakami (audio engineers), Robert Pahle (IT support), Pasan Dharmasena (rendering of avatars). We would like to thank also the Future Reality Lab at NYU for lending the space and technology.

## 6. References

- [1] Kris Layng, Ken Perlin, Sebastian Herscher, Corinne Brenner, and Thomas Meduri, “Cave: Making collective virtual narrative: Best paper award,” *Leonardo*, vol. 52, no. 4, pp. 349–356, 2019.
- [2] David Gochfeld, Corinne Brenner, Kris Layng, Sebastian Herscher, Connor DeFanti, Marta Olko, David Shinn, Stephanie Riggs, Clara Fernandez-Vara, and Ken Perlin, “Holojam in wonderland: Immersive mixed reality theater,” *Leonardo*, vol. 51, no. 4, pp. 362–367, 2018.
- [3] Thomas Deacon, Tony Stockman, and Mathieu Barthet, “User experience in an interactive music virtual reality system: An exploratory study,” in *Bridging People and Sound*, Mitsuko Aramaki, Richard Kronland-Martinet, and Sølvi Ystad, Eds., Cham, 2017, pp. 192–216, Springer International Publishing.
- [4] Stefania Serafin, Cumhuri Erkut, Juraj Kojs, Niels C. Nilsson, and Rolf Nordahl, “Virtual reality musical instruments: State of the art, design principles, and future directions,” *Computer Music Journal*, vol. 40, no. 3, pp. 22–40, 2016.
- [5] I. Bergström, S. Azevedo, P. Papiotis, N. Saldanha, and M. Slater, “The plausibility of a string quartet performance in virtual reality,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 4, pp. 1352–1359, April 2017.
- [6] Byungdae Jung, Jaemin Hwang, Sangyoon Lee, Gerard Jounghyun Kim, and Hyunbin Kim, “Incorporating co-presence in distributed virtual music environment,” in *Proceedings of the ACM symposium on Virtual reality software and technology*. ACM, 2000, pp. 206–211.
- [7] Pontus Larsson, Aleksander Våljamäe, Daniel Västfjäll, Ana Tajadura-jiménez, and Mendel Kleiner, “Auditory-Induced Presence in Mixed Reality Environments and Related Technology,” in *The Engineering of Mixed Reality Systems*, chapter 8, pp. 143–163. Springer-Verlag, 2009.
- [8] Ina Wagner, Rod Mccall, Ann Morrison, and Marne Valle, “On the Role of Presence in Mixed Reality,” *Presence*, vol. 18, no. 4, pp. 249–276, 2009.
- [9] Saniye Tugba Bulu, “Place presence, social presence, co-presence, and satisfaction in virtual worlds,” *Computers & Education*, vol. 58, no. 1, pp. 154–161, 2012.
- [10] Will Bailey and Bruno M. Fazenda, “The effect of visual cues and binaural rendering method on plausibility in virtual environments,” in *Proceedings of the 144th AES Convention*, Milan, Italy, 2018.
- [11] Durand R. Begault, Elizabeth M. Wenzel, and Mark R. Anderson, “Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source,” *Journal of the Audio Engineering Society. Audio Engineering Society*, vol. 49, no. 10, pp. 904–916, 2001.
- [12] Stefan Liebich, Raphael Brandis, Johannes Fabry, Peter Jax, and Peter Vary, “Active occlusion cancellation with hear-through equalization for headphones,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 241–245.
- [13] Daniel L Valente and Jonas Braasch, “Subjective scaling of spatial room acoustic parameters influenced by visual environmental cues,” *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 1952–1964, 2010.
- [14] Jens Blauert, *Spatial hearing: the psychophysics of human sound localization*, MIT press, 1997.
- [15] “Motive - optical motion capture software,” <https://optitrack.com/products/motive/>, (Accessed on 08/27/2019).
- [16] “Unity real-time development platform — 3d, 2d vr & ar visualizations,” <https://unity.com/>, (Accessed on 08/01/2019).
- [17] “Autodesk maya — 3d computer animation, modeling, simulation, and rendering software,” <https://www.autodesk.com/products/maya/overview>, (Accessed on 08/27/2019).
- [18] Julian Vanasse, Andrea Genovese, and Agnieszka Roginska, “Multichannel impulse response measurements in matlab: An update on scanir,” in *Proceedings of the AES International Conference on Immersive and Interactive Audio*, York, UK, 2019.
- [19] Angelo Farina, “Simultaneous measurement of impulse response and distortion with a swept-sine technique,” in *Audio Engineering Society Convention 108*. Audio Engineering Society, 2000.
- [20] “Steam audio,” <https://valvesoftware.github.io/steam-audio/>, (Accessed on 08/01/2019).
- [21] BG Shinn-Cunningham, “Distance cues for virtual auditory space,” *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia*, pp. 227–230, 2000.
- [22] Cristina Rottondi, Chris Chafe, Claudio Allocchio, and Augusto Sarti, “An overview on networked music performance technologies,” *IEEE Access*, vol. 4, pp. 8823–8843, 2016.



# Abstract Reviewed Paper at ICSA 2019

Presented \* by VDT.

## Data set: BRIRs for position-dynamic binaural synthesis measured in two rooms

A. Neidhardt<sup>1</sup>

<sup>1</sup> *Technische Universität Ilmenau, Germany, Email: annika.neidhardt@tu-ilmenau.de*

### Abstract

Binaural room impulse responses were measured with a KEMAR 45BA head-and-torso-simulator. For the first data set, it was placed at different positions located on a line with a length of 2 m in a 25 cm positional resolution and an azimuth resolution of 4°. Two source positions were considered in the setup, one in front of the line, one at the side. The same arrangement of source and receiver positions was realized in two different rooms, a quite dry listening laboratory and a quite reverberant seminar room. For the second data set, BRIRs and omni-directional RIRs were measured for a translation line with a length of 7.5 m through the given seminar room. The data sets are valuable for realizing, testing and studying dynamic binaural walk-through scenarios in the two different rooms.

## 1. Introduction

Today, devices for experiencing virtual and augmented reality are available to everyone and can be used even at home. The user can move within a certain area that is covered by motion capturing modules. For devices like head mounted displays the audio reproduction is usually realized over headphones with dynamic binaural synthesis.

To provide efficient audio rendering algorithms for position changes of the listener, researchers are currently investigating potential for simplification in psychoacoustical and data-driven studies. Various approaches for interpolation and extrapolation are considered. In this context data sets with binaural room impulse responses (BRIRs) measured at densely arranged positions for varying head rotation angles are of interest, e.g. to provide a reference scenario.

Within this publication, the measurement of such data sets for two different rooms is documented. The created data sets are provided for free download.

## 2. Data set 1 - 2 m line in two rooms

The same arrangement of loudspeakers and a line of 9 listening positions with a length of 2 m was realized in two rooms, a relatively dry listening laboratory and a quite reverberant seminar room. The direct sound is similar in both scenarios, but the amount of reverberant energy as well as the spatio-temporal structure of the early reflections differ with the room.

### 2.1. Room 1 - Listening laboratory

The first room is the listening laboratory of the university in Ilmenau. The room has a size of 8.4 m × 7.6 m × 2.8 m, a volume of  $V = 179 \text{ m}^3$  and a reverberation time  $T_{60} = 0.27 \text{ s}$  (broad band). It complies recommendation ITU-R BS.1116-2.

A translation line with a length of 2 m was defined within the room as illustrated in Figure 2. A G.R.A.S. KEMAR 45BA head-and-torso-simulator (HATS) with large ears was positioned on an electronic turntable Outline ET 250-3D for accurate rotation. Two loudspeakers Genelec 1030A were placed in the room, one in front of the translation line, the other at the side as illustrated in Figure 2.



Fig. 1: The measurement setup in the listening laboratory.

The translation line covers the distances of 1.25 m to 3.25 m to the center of the frontal loudspeaker and passes the second loudspeaker with a minimum distance of 1.25 cm (center head to center loudspeaker).

BRIRs were measured at nine positions with equal distances of 25 cm along the translation line. At each of the positions measurements were conducted with an azimuth resolution of 4° over the full 360° rotation of the HATS. Elevation changes were not considered. For the measurement a swept sine method with a logarithmic sweep ranging from 50 Hz to 20 kHz over a duration of 3 s was used.

### 2.1.1. Psychoacoustic evaluation and further experiments

In order to conduct a psychoacoustic evaluation, Neidhardt and Knoop [1] asked subjects to rate the plausibility of the position-dynamic binaural reproduction in a Yes/No paradigm without including a real version of the scene. With regard to their inner reference all participants rated the BRIR data set as plausible for the frontal loudspeaker reproducing male speech and a pop song.

In another study Neidhardt et al. [2] investigated the perceptual consequences of systematic simplifications of the BRIR data set in an interactive position-dynamic exploration scenario. The original data set was included as one of the test cases and was again rated as plausible by all participants.

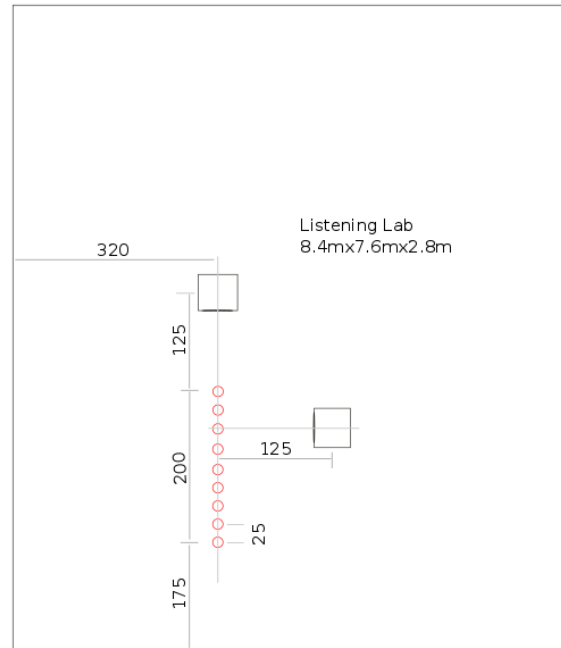


Fig. 2: Measurement positions of the Kemar 45BA and the two loudspeakers in the listening laboratory.

In both studies STAX SR-202 headphones with a non-individual headphone compensation filter with roll-off frequencies of 80 Hz and 20 kHz were used. The headphone filter was created from a measurement with STAX headphones positioned on the Kemar 45BA by the least squares approach according to the description by Schärer and Lindau [3].

Furthermore, in both studies pyBinSim [4] was used for the dynamic auralization. When switching between filters only a very short cross-fade in the time domain was applied. No interpolation or extrapolation was used.

## 2.2. Room 2 - Seminar room

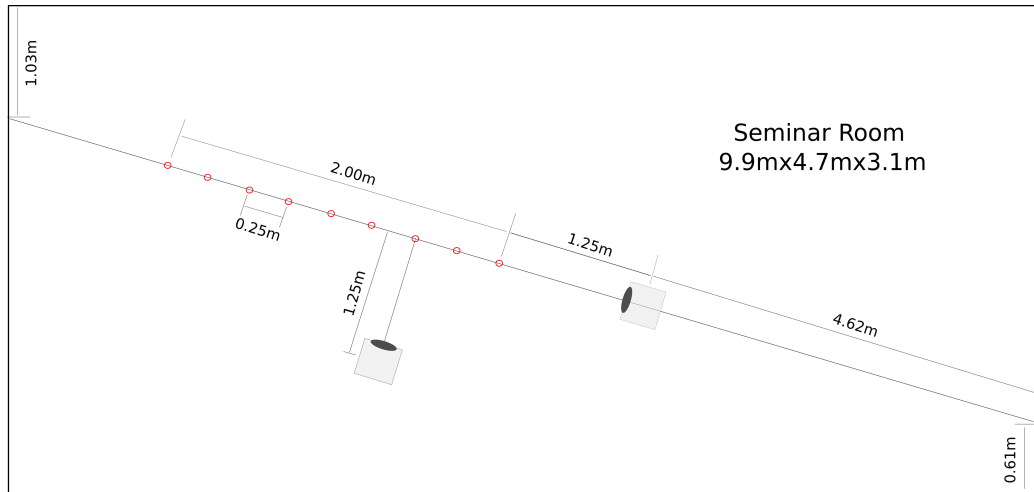
The same arrangement of source and receiver positions was set up in a seminar room of the university. The room has a size of 9.9 m × 4.7 m × 3.1 m, a volume of  $V=144\text{ m}^3$  and a reverberation time of  $T_{60}=0.99\text{ s}$  (broadband).

Again, the HATS was placed on the turntable at 9 positions along the translation line with the length of 2 m. The location of the translation line in the room is illustrated in Figure 3. A photo of the measurement setup shown in Figure 4. BRIRs were measured with an azimuth resolution of 4°.

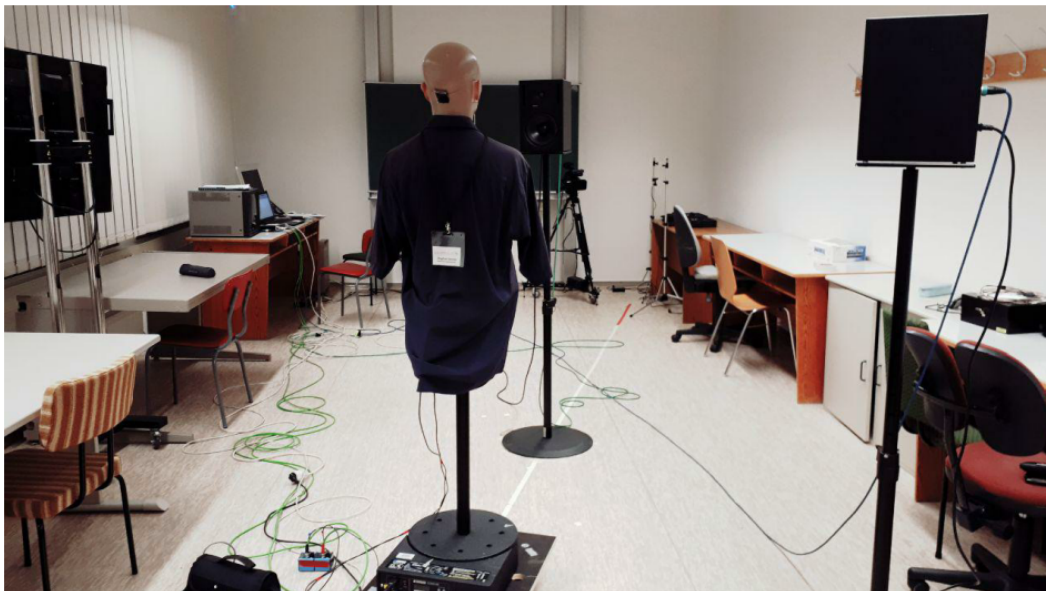
### 2.2.1. Additional BRIRs: Same setup with headphones

In order to study the plausibility and authenticity as suggested in [5] and [6], BRIRs need to be measured for a scenario that can be directly compared to a real sound field. In such test subjects should not need to take the headphones off and put it on again. To create BRIRs for such a test paradigm, open headphones were placed on the HATS as shown in Figure 5.

In previous studies, extra-aural headphones were used. The



**Fig. 3:** Measurement positions of the Kemar 45BA and the loudspeakers Genelec 1030A in the seminar room.



**Fig. 4:** Measurement setup in the seminar room.

data set presented in this study should enable the listener to move free and intuitively along the given line. The extra-aural headphones are quite heavy and instable on the head. Listeners usually carry out only slow careful movements while wearing those. As a consequence, we decided to use the open AKG K1000 headphones instead. They were placed on the dummy, which was moved only very carefully to make sure the headphones did not change their position on the head during the whole measurement.

Furthermore, the headphone transfer function (HpTF) was measured right away with the same positioning on the head. The headphone compensation filter created from this measurement is included in the data set.

For an authenticity study an azimuth resolution of  $4^\circ$  may not be enough. Therefore this additional BRIR set was measured with an azimuth resolution of  $2^\circ$ .

**2.2.2. Psychoacoustic evaluation and further experiments**

The BRIRs measured without the headphones on the HATS were used in a study similar to [2] with the goal to verify the observations in a more reverberant room. The results have not been published yet. Also in this study the measured data set was rated as plausible by the participants in a test paradigm that did not include a real version of the sound field. Continuity, externalization and the impression of walking towards a sound source were rated very well, too. The experiment included only the BRIRs for the frontal loudspeaker and the test was only conducted with dry male speech as a test signal. The results of [2] indicate that for noise the positional resolution is not adequate.

A psychoacoustic experiment [7] was conducted to evaluate plausibility in the Yes/No paradigm as suggested by Lindau and Weinzierl [5]. A detailed documentation of the experiment and the results, observations and discussion





**Fig. 5:** Headphones AKG K1000 with a  $45^\circ$  opening angle were placed on the HATS in order to enable a direct comparison of the binaural reproduction with the real loudspeaker reproduction.

can be found in the corresponding paper [7]. Without an individualization, experts were able to identify the binaural auralization, though they found it challenging in several cases. Most of the untrained listeners did not find reliable cues for an identification.

### 3. Data set 2 - 7.5 m in seminar room

In order to investigate the position dependent change of BRIRs over a longer distance, a second BRIR data set was created in the same seminar room. A translation line with a length of 7.50 m was defined diagonal through the rooms. The line is an extension of the previous one, several listening positions are equivalent. However, the source positions were changed according to the illustration in figure 6. For this data set, the Kemar 45BA was placed at 16 positions along the line with a 50 cm spacing in between. The measurement positions cover a distance from 1 m to 8.50 m to the frontal loudspeaker. Again, for each of the positions and both loudspeakers BRIRs were measured with an azimuth resolution of  $4^\circ$  for the full  $360^\circ$  rotation.

#### 3.1. Turning the loudspeakers by $180^\circ$

The directivity of a sound source influences the sound field and the acoustical reflection pattern in a room. To study this effect, the loudspeakers at both positions were turned by  $180^\circ$  around their center. Consequently, the loudspeaker membrane was moved to the other side. The whole series of measurements was repeated. The setup is shown in figure 7.

#### 3.2. Omnidirectional RIRs

For a more detailed analysis, an omni-directional Microtech Gefell MK221 microphone capsule connected to a MV203 amplifier unit was placed subsequently at the center of each of dummy head positions. The same logarithmic sweep was used to measure the omni-directional room impulse responses at each of the positions.

## 4. Conclusion

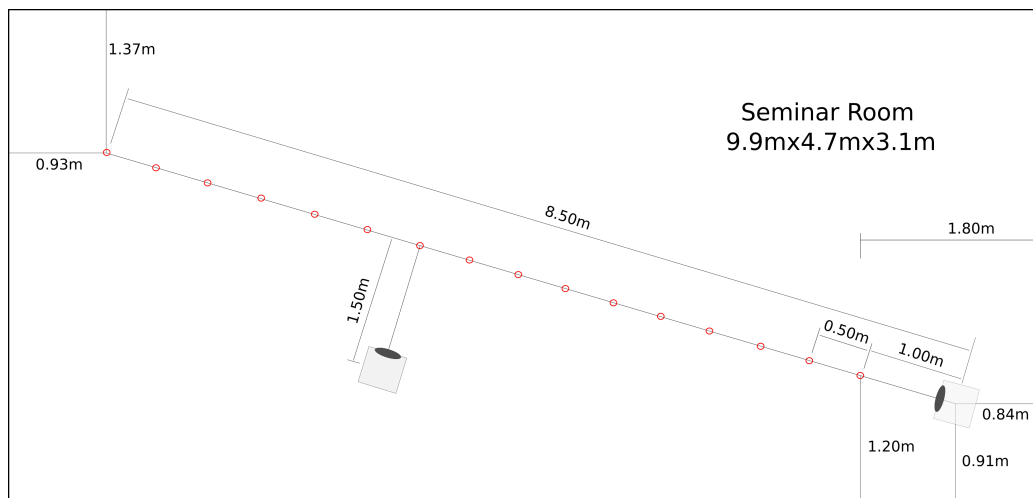
This paper documents the creation of BRIR data sets for the realization of position-dynamic binaural synthesis. A comparable measurement setup was realized in two different rooms, a relatively dry listening laboratory and a quite reverberant seminar room. The created data sets are available as creative commons CC BY-SA 4.0 following the links provided in [8] and [9].

## 5. Acknowledgement

Thanks to Anna-Maria Zerlik, Samaneh Kamandi and Anson Davis Pereppadan for their contributions to the data sets. This study was funded by the DFG (German Research Council, BR 1333/18-1).

## 6. References

- [1] A. Neidhardt, N. Knoop: Binaural walk-through scenarios with actual self-walking using an HTC Vive. 43rd Annual Conference on Acoustics, Kiel, Germany, 2017.
- [2] A. Neidhardt, A.I. Tommy, A.D. Pereppadan: Plausibility of an interactive approaching motion towards a virtual sound source based on simplified BRIR sets. 144th Int. AES Convention, Milan, Italy, 2018
- [3] Z. Schärer, A. Lindau: Evaluation of Equalization methods for Binaural signals. 126th Int. AES Convention, Munich, Germany, 2009
- [4] A. Neidhardt, F. Klein, N. Knoop and T. Köllmer: Flexible Python tool for dynamic binaural synthesis applications. 142nd Int. AES Convention, Berlin, Germany, 2017.
- [5] A. Lindau, S. Weinzierl: Assessing the plausibility of virtual acoustic environments. Acta Acustica united with Acustica 98, 804-810, 2012.
- [6] F. Brinkmann, A. Lindau, S. Weinzierl: On the authenticity of individual dynamic binaural synthesis. The Journal of the Acoustical Society of America 142, 1784 (2017); doi: 10.1121/1.5005606
- [7] A. Neidhardt, A.-M. Zerlik: Plausibility versus authenticity of binaural walk-through scenarios. 5th International Conference on Spatial Audio (ICSA), Ilmenau, Germany, 2019
- [8] A. Neidhardt, A.-M. Zerlik, S. Kamandi: Binaural room impulse responses for interactive listener translation in two rooms. (Version V1.0) [Data set]. Zenodo (2019). <http://doi.org/10.5281/zenodo.3457782>
- [9] A. Neidhardt, A.D. Pereppadan: Binaural room impulse responses for interactive listener translation in a seminar room. (Version V1.0) [Data set]. Zenodo (2019). <http://doi.org/10.5281/zenodo.3457798>



**Fig. 6:** Measurement positions of the Kemar 45BA and the loudspeakers Genelec 1030A for a translation line throughout the seminar room.



**Fig. 7:** Measurement setup in the seminar room with the loudspeakers turned by 180°.







## Abstract Reviewed Paper at ICSA 2019

Presented by VDT.

### 3D audio for live events

L. Nipkow

*Silent Work GmbH, Zurich, Switzerland, Email: lasse.nipkow@silentwork.com*

#### Abstract

Many applications for 3D audio are aimed at live events. Several manufacturers of PA systems now offer solutions. These are not optimized for native 3D audio, but achieve an immersive impression due to placement of objects and added reverberation.

To operate native 3D audio for live events, there are two key questions to consider:

- How should the loudspeaker layout be chosen so that the audience gets the most impressive listening experience at a reasonable cost?
- How should 3D audio content be designed so that it reaches the audience as effectively as possible through such PA systems?

Two phenomena can be used for 3D audio: Envelopment and projection of sound sources in the front. The most impressive case is provided by a loudspeaker setup using imaginary connecting lines between the loudspeakers, which results in a volume. Vertical phantom sound sources presented in the front sound more natural than horizontal ones. The number of vertical front loudspeakers determines the resolution of the image.

The wedge-shaped loudspeaker setup, that means with height loudspeakers in the front, meets both requirements minimally. The greater the listening area, the more loudspeakers are needed. The greater the distances of the loudspeakers to each other, the more audible holes are perceived, depending on the acoustics of the playback room.

# 1. Introduction

3D audio has enormous potential: listeners can experience previously unheard sound at events. This is especially the case, as long as 3D audio setups are rarely found in homes, cinemas and cars and thus consumers do not have access to it.

To make the most out of 3D audio, event organizers must meet some requirements; these are very complex:

- Guarantee optimal room acoustics.
- Provide a sufficiently high density of loudspeakers in the reproduction room.
- Set up a meaningful loudspeaker layout so that psycho-acoustical effects such as vertical stereophony and the cocktail party effect can be used.
- Produce suitable content that uses the loudspeaker setup highly natively.

Depending on the situation, such as open-air situation or addressing a hall, requires different measures to achieve a convincing result.

## 2. Loudspeaker layout and psycho-acoustics

### 2.1. Initial situation for the playback of 3D audio

Sound engineers create reproducible 3D audio content primarily in studio and movie rooms. These rooms sound very good in contrast to large event halls. This raises the question of how to best scale up mobile PA systems for 3D audio, in order to lose the least noticeable added value of 3D audio. Therefore, it is crucial to look at the appearing phenomena of 3D audio and to investigate which perceptible sound changes occur when the scaling factor changes. For this purpose, the loudspeaker layout of the two components sound sources (direct sound) and envelopment (room sound) are first considered in isolation from each other and then combined.

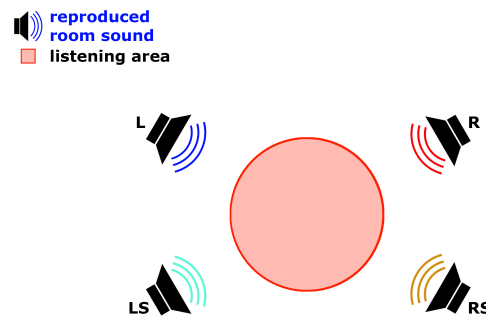
### 2.2. Loudspeaker layout for envelopment

Envelopment is one of the most important features of 3D audio. Spatiality comes from sound, which is reflected in rooms at the boundary surfaces; the sound of the room meets the listener distributed equally from all directions. Convincingly appearing envelopment therefore requires addressing of the listener with room sound from different directions.

#### 2.2.1. Envelopment for the horizontal plane

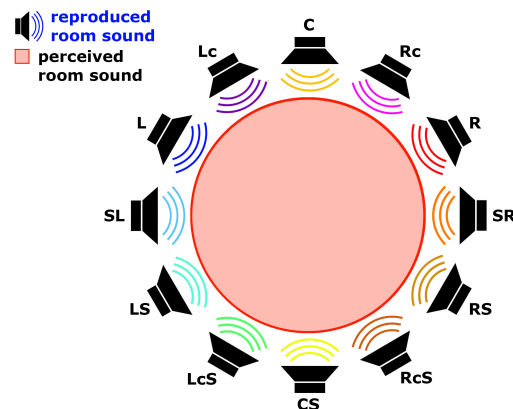
For the simplest case of an envelopment by PA, four loudspeakers are used; they stand at the same distance around the listener on the horizontal plane, see Fig. 1.

This works very convincingly in the sweet spot for listening situations in control rooms and living rooms. However, if the volume of the room, and thus the area of the loudspeaker setup is much larger, i.e. this means more than 5m between



**Fig. 1:** View from above on the loudspeaker setup: The listener already feels impressively enveloped by four loudspeakers that reproduce the room sound. This requires relatively small distances between the loudspeakers (up to about 5m x 5m). Due to the small number of loudspeakers, the resulting listening area is small. When a listener approaches one of the loudspeakers, the level rises from that direction. If the level exceeds 10 dB, the remaining loudspeakers are no longer effective for an envelopment.

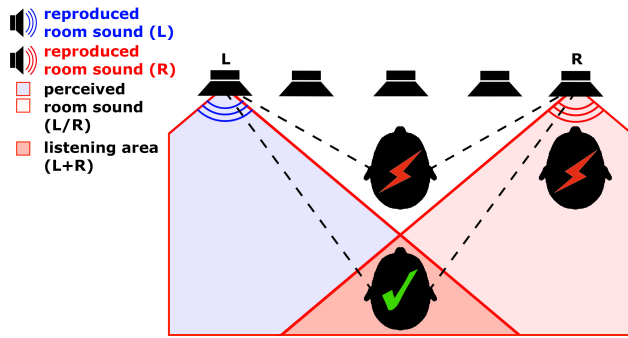
adjacent loudspeakers, noticeable holes between the loudspeakers occur outside the sweet spot. This can be avoided by placing the loudspeakers in a sufficiently high resolution around the listener; thus the mentioned perceptible holes do not appear, see Fig. 2.



**Fig. 2:** View from above onto the loudspeaker setup: When loudspeakers reproduce room sound around the listener from all horizontal directions, the listener feels a very impressive envelopment. Since many loudspeakers reproduce room sound from all directions, the room sound is distributed very regularly in the room. A listener can get much closer to a loudspeaker before its sound exceeds the 10dB threshold.

On the one hand, the cost of a high-resolution loudspeaker setup increases with each additional loudspeaker. On the other hand, the resolution also has an influence on the perceptibility of the individual loudspeakers as sound sources. The room sound is very balanced in the center of the listening area. This means from that position, with a favorable signal selection each loudspeaker signal is perceptible. The more a listener approaches a loudspeaker, the louder its signal becomes in relation to the other loudspeakers. If, at the location of the listener, the level of the nearest loudspeaker exceeds 10 dB compared to other loud-

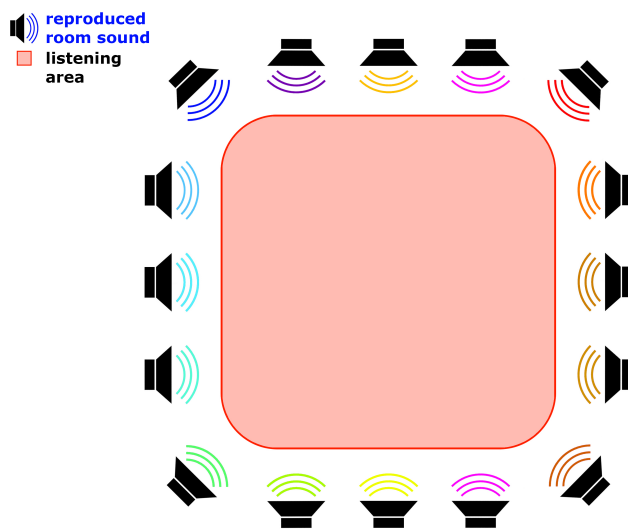
speaker signals, a clear misbalance of the spatiality occurs [1]; the listener perceives only the loudest loudspeaker, compare cocktail party effect [2], see Fig. 3.



**Fig. 3:** View from above on a part of a 3D audio loudspeaker setup: If a listener is in the listening area, he / she hears both room sound signals from left and right about equally loud and at a relatively small angle. If the person is close to a loudspeaker, he / she practically only hears that signal. If the listener is very close to the front between two loudspeakers, the angle between the two loudspeakers becomes very large: In this case, the listener perceives a hole in the middle.

The optimal resolution or optimal distance between the adjacent loudspeakers can be determined on the basis of the radiation characteristic (dispersion): The narrower the dispersion angle, the closer the loudspeakers must be to one another. Furthermore, the minimum possible distance between a person and the loudspeakers can be influenced: with a doubling of the resolution, the minimum possible distance is halved.

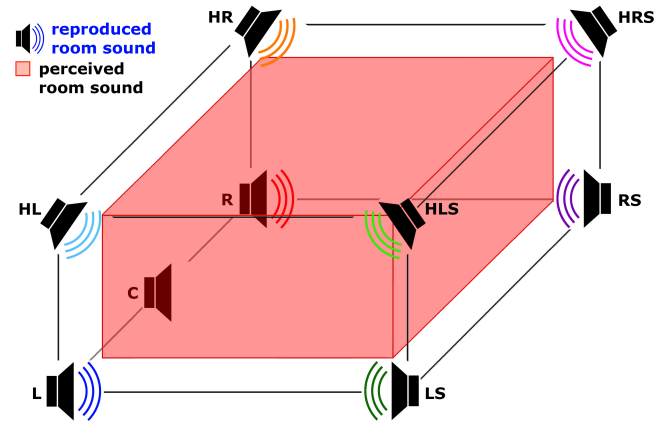
The expense of up-scaling such a loudspeaker arrangement and at the same time avoiding perceptible holes becomes greater as the listening area increases. Fig. 4 shows a case of a square loudspeaker setup with a width of 20m or a listening area of about 400m<sup>2</sup>.



**Fig. 4:** View from above onto a large loudspeaker setup with a listening area of about 400m<sup>2</sup>: The loudspeaker ring produces a homogeneous sound field with room sound at a sufficiently high resolution.

### 2.2.2. Envelopment for 3D audio

The effect of the envelopment becomes more impressive when the loudspeaker setup spans a volume rather than a horizontal plane. In this case, the listener has the impression of being in the recording room when using meaningfully selected loudspeaker signals, see Fig. 5.



**Fig. 5:** The reproduced room sound between the front and rear loudspeakers should be balanced in level to ensure an impressive envelopment. In this case, a listener can be placed almost anywhere in the setup and has the impression of being in the recording room.

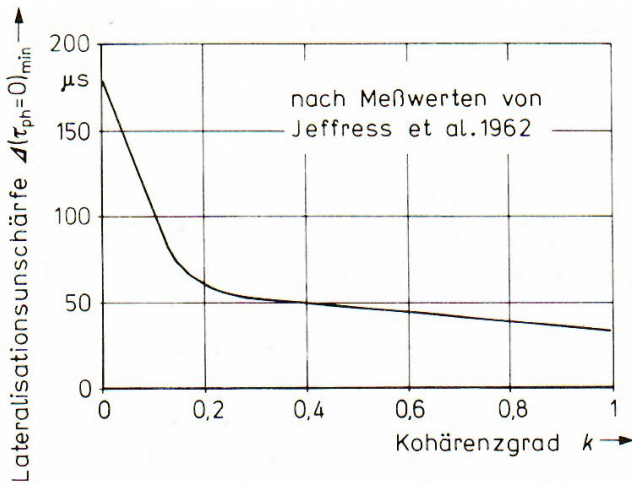
The expense of up-scaling such a loudspeaker arrangement and at the same time avoiding perceptible holes increases with a larger listening area as well here. The following chapters show that due to psychoacoustic laws, a loudspeaker arrangement requires much fewer loudspeakers than expected from a geometrical point of view.

### 2.3. Signal quality for optimal envelopment

Perceived envelopment occurs during loudspeaker reproduction, when all loudspeakers in the setup reproduce room signals of similar levels and a correlation around zero [1]. The correlation has a considerable influence on the perceived spatiality; it increases prominent at a degree of coherence of less than  $k = 0.2$  (ear signals), see Fig. 6 [2].

The degree of coherence of the sound source signals is usually different from that of the ear signals. Incoherent sound source signals result in incomplete incoherent, i.e. in partially coherent ear signals [2]. This means that the correlation of loudspeaker signals must be less than the degree of coherence in ear signals in order to produce the maximum possible spatiality; the optimum is around zero.

A correlation of 1, however, is very unfavorable for room signals; correlation 1 corresponds to a sound with identical loudspeaker signals. There may be strong audible comb filtering effects. Spatiality, on the other hand, occurs primarily in reflected sound from side walls [3].



**Fig. 6:** Lateral blurring  $\Delta$  ( $\tau_{ph}=0$ )min as a function of the degree of coherence of the ear signals. Low pass noise  $f_g = 2\text{kHz}$ , level about 90dB, 7 subjects. The measured value for  $k = 0$  corresponds to complete uncertainty in the experimental setup used [2].

Sound engineers should avoid such signals for the side loudspeakers in particular; in many cinemas with a surround sound system, however, this is exactly what is done by distributing few signals by the audio playback systems to many surround loudspeakers.

## 2.4. Reduction of the number of loudspeakers

Especially in a mobile PA system, the outlay must be in a sensible relationship to the benefit. Therefore, loudspeakers which provide no significant added value in the set position in the loudspeaker setup should be omitted.

### 2.4.1. Reduction of height loudspeakers

At many live events there is a stage or a clear orientation of the audience to the front. This means that side loudspeakers lead to a two-sided lateral addressing of the audience with room sound.

Lateral sound addressing on both sides of the listener leads in turn to emphasis of high frequencies [4] among others also in the range around 8kHz similarly to height loudspeakers of a 9.1 loudspeaker setup. This means that the directional bands elevate room sound with a correlation around zero, see Fig. 7.

Since these are signals with a correlation around zero and not identical signals, the elevation effect between bottom and top is blurred; this contributes to the filling of the above-mentioned holes. The balance between the left and right sides may be around 15dB when used with a side-positioned loudspeaker; at higher level differences, the elevation effect decreases rapidly, and the listener can only perceive the sound of the louder loudspeaker as a real sound source.

Due to comb filtering effects lateral addressing also leads to elevation effects with lower intensity even at low frequencies at 600 Hz [5], see Fig. 7.

### 2.4.2. Reduction of the rear loudspeakers

Listeners who are located far in front of the listening area can not perceive room sound from behind because of the masking due to direct sound and lateral addressing due to room sound; the level difference is much higher than 10dB, compare cocktail party effect [2].

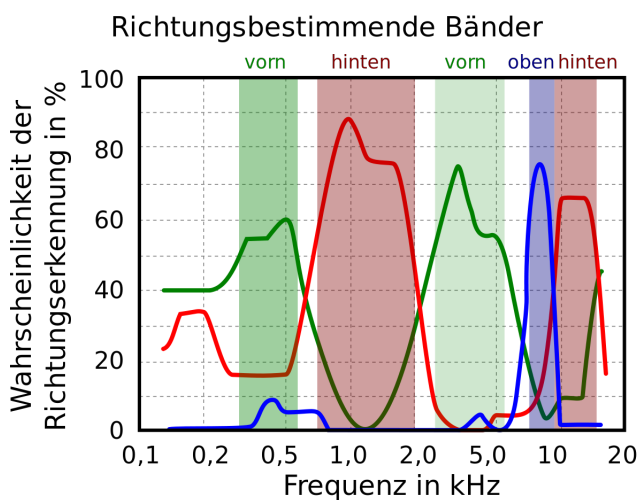
As described above, both-sided lateral addressing of listeners leads to an emphasis of high frequencies. In addition to 8kHz higher frequencies around 10kHz are raised and lead to a perception from behind. Thus, listeners have the impression of perceiving reverberation also from the top and rear when using side loudspeakers, see Fig. 7.

An addressing from the rear is associated with risks: Listeners who are located far back in the listening area and thus near the rear loudspeakers, primarily hear the sound of those loudspeakers. As a result, sound from all other directions, and especially from the front, may be heavily masked.

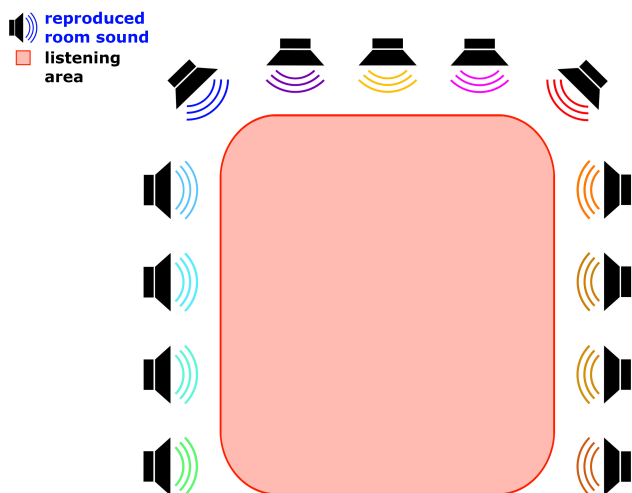
Room sound from behind leads to a narrowing of the stereo panorama for room sound and an increase in the degree of coherence of the ear signals respectively. Therefore, such signals do not contribute to the spatiality but only to an avoidance of perceptible holes.

Furthermore, concertgoers in halls or churches and in the vicinity of the back wall do not perceive room sound from that direction. Due to their individual position in the room, they perceive room sound primarily from the direction where the largest proportion of the room volume lies.

All of these reasons argue that loudspeakers should not be used on the back of the loudspeaker setup regarding room sound. This results in the following structure for the setup, see Fig. 8.



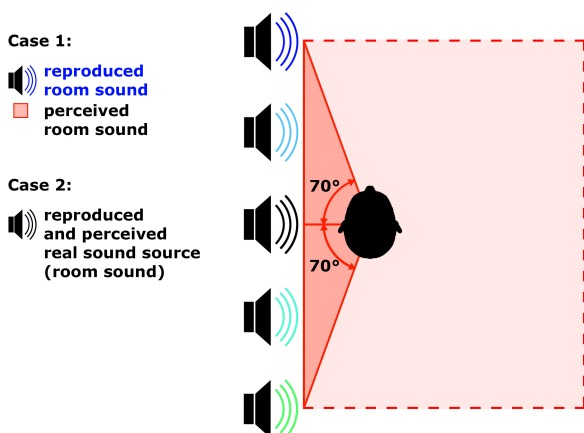
**Fig. 7:** The directional bands [2] make it possible to omit loudspeakers in the setup without compromising on sonority: Listeners perceive both sides of laterally reproduced room sound with an increase in high frequencies. As a result, the room sound not only appears from the sides but also from both above and from behind.



**Fig. 8:** View from above of a loudspeaker setup for a 3D audio PA system without height loudspeaker as well as without rear loudspeakers. Due to psychoacoustic phenomena, the room sound appears both from the back and from above for listeners within the loudspeaker setup at a favorable signal selection.

**2.4.3. One-sided spatiality**

One-sided spatiality occurs when room sound reaches the listener only from one side. If, instead of a real sound source, several loudspeakers are arranged on a line to the side of the listener, the reproduced sound covers an angle of sound incidence of up to approximately  $\pm 70^\circ$ . Since in this case several loudspeakers reproduce sound with correlation around zero, the listener does not perceive a real sound source anymore, but sound that largely corresponds to side reflections. This type of sound addressing also leads to elevation effects. At least 4 loudspeakers are necessary for this, so that on the one hand the angle of sound incidence is sufficiently large and on the other hand no perceptible holes occur between the lateral loudspeakers, see **Fig. 9**.



**Fig. 9:** View from above onto a loudspeaker setup. In case 1, 5 loudspeakers reproduce room sound with a correlation around zero. The listener perceives a one-sided spatiality on the left, extending from the front to the rearmost loudspeaker. This corresponds to slightly less than half (about 40%) of an envelopment (100%) on the horizontal plane. In case 2 only one loudspeaker provides the listener from the side with sound. The listener perceives a real sound source instead of a spatiality; this corresponds to a point on the horizontal plane.

**2.5. Loudspeaker layout for sound sources**

For the simplest case of imaging by sound addressing a stereo configuration is used. On a wide stage with 10m or even 20m, the resulting stereo base width is too large; it would create a middle hole. In addition, in the case of phantom sources outside of the middle axis, some serious distortions of the image arise: the farther a listener moves away from the middle axis, the further the image shifts in the direction of the closer loudspeaker [6]. Therefore, the solution is similar to room sound: a sufficiently high number of loudspeakers within the stereo base.

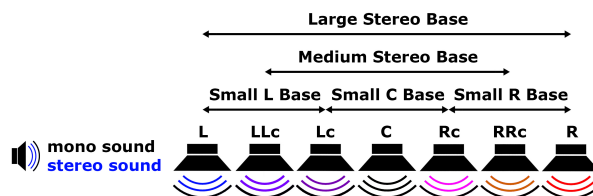
**2.5.1. Combination of real sound and phantom sources**

An extension of a 2-channel stereo setup with additional loudspeakers raises the question as to what criteria the sound engineer should use to assign signals to the loudspeakers.

In the simplest case, the loudspeakers within the stereo base represent real sound sources. The direct sound of mono-captured or strongly directed instruments such as vocals, trumpet, bass drum, e-bass, etc. has no stereophonic width. This sound content is therefore predestined to be used as real sound sources. In addition, the transparency of real sound sources is much greater than that of phantom sources [6]. This has a dramatic effect, especially at high frequencies.

Acoustic instruments with resonating bodies such as grand pianos, strings and large bodies of sound such as choirs that are captured stereophonically require a stereo width in the reproduction in order to sound as natural as possible.

In general, the wider the stereo base for stereophonic microphones is, the more impressive the horizontal stereo reproduction becomes. Conversely, the narrower the chosen stereo base is, the more powerful / compact the sound sources will be. If the number of loudspeakers in the front is sufficiently high, stereo bases of different widths can be created and also several narrow stereo bases next to each other, see **Fig. 10**. Signals from stereo microphones and synthesizers or similar sources can be used. When using different adjacent narrow stereo bases, the instruments spatially delimit clearer from each other without sounding mono.



**Fig. 10:** View from above onto the loudspeaker setup: If there are enough loudspeakers in the front, different widths of stereo bases can be created and also several narrow stereo bases next to each other. Each loudspeaker can represent a real sound source, see center loudspeaker ,C'.



In some situations, it makes sense to combine phantom and real sound sources for instruments. A typical case is the bass drum: The attack should appear in the center channel and the boost distributed over several loudspeakers in order to achieve a spatial extent of the instrument.

### 2.5.2. Projection of sound sources in the front

In the context of direct sound, early reflections lead to a distance impression of sound sources [7]. This occurs when direct sound and early reflections are reproduced from the same direction. If the lower loudspeakers in the front L-C-R reproduce direct sound predominantly and the upper loudspeakers in the front HL-HR reproduce early reflections, this results in an audible connection of the lower and upper front levels; this is what the author describes as a projection of the sound body in the front [8], see Fig. 11.

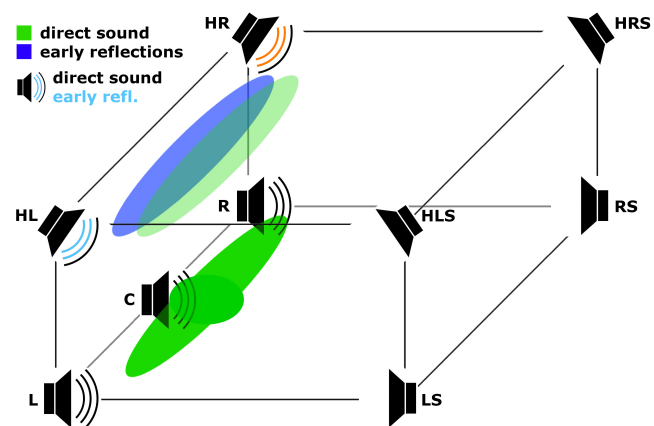


Fig. 11: If the lower loudspeakers in the front L-C-R reproduce direct sound predominantly and the upper loudspeakers in the front HL-HR reproduce early reflections, this results in an audible connection of the lower and upper front levels: projection in the front.

The instruments are mainly located at the lower loudspeakers due to direct sound. However, they sound more natural than without the early reflections; it corresponds approximately to the conditions in a concert hall near the stage.

### 2.5.3. Vertically reproduced direct sound

If one of the lower loudspeakers and the loudspeaker vertically above it reproduce stereophonic direct sound with sound content correlation [8], such as L-HL, this results in a similar natural effect as in early reflections, see Fig. 12.

In this case, the image is not clearly perceived from the lower loudspeaker but is located between the two loudspeakers involved. Noteworthy in this case is the localization sharpness in the horizontal plane, which corresponds to that of a real sound source [6]. The combination of a very sharp localization with a simultaneously strong naturalness of the instrument and impressiveness of a synthetic sound means that the sound source is perceived very direct and therefore attracts attention. This is especially true for sound sources with a high proportion of high frequencies [8].

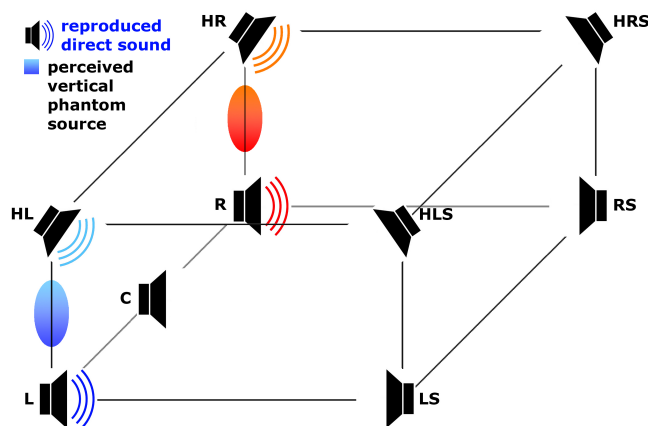


Fig. 12: In this 9.1 loudspeaker setup, two stereo pairs of two vertical instruments on the front edge reproduce two instruments / sounds. The transparency of a mix remains the same as if real sound sources were used for a surround sound reproduction.

The lower and upper loudspeakers must sound similar and have a wide dispersion characteristic both vertically and horizontally for vertical stereophony to work reliably. Only in this way it is possible for a listener to perceive both loudspeaker signals equally loud and, on the other hand, balanced levels between several vertical stereo pairs, largely independent of their position in the reproduction room.

Furthermore, vertical stereophony works optimally only if the lower and upper loudspeakers are perpendicular to each other and have no horizontal components. Otherwise, partially horizontal phantom sound sources arise.

As described above, vertical stereo pairs in the front edges of the loudspeaker setup are not enough at a 10m or 20m stereo base; there are noticeable holes in the image. The solution lies in a loudspeaker layout with two identical height levels of loudspeakers. Thus, both the above-described requirements for vertical arrangement and high resolution are simultaneously fulfilled, see Fig. 13.

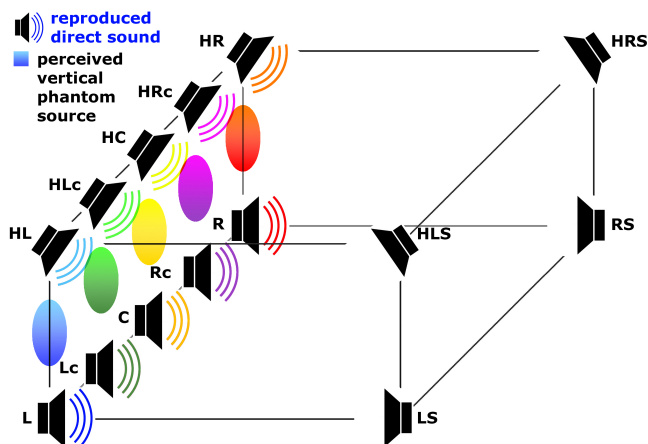


Fig. 13: In this extended 9.1 loudspeaker setup, five vertically and equidistantly arranged stereo pairs in the front represent instruments / sounds. The transparency is very high in this constellation compared to horizontal phantom sources in the front.

## 2.6. Loudspeaker layout for sound sources and envelopment

Sound sources (direct sound) and envelopment (room sound components) always occur at the same time and overlap each other in a PA system. Therefore, the result for a PA system with 3D audio is the combination of the two partial results, see Fig. 14.

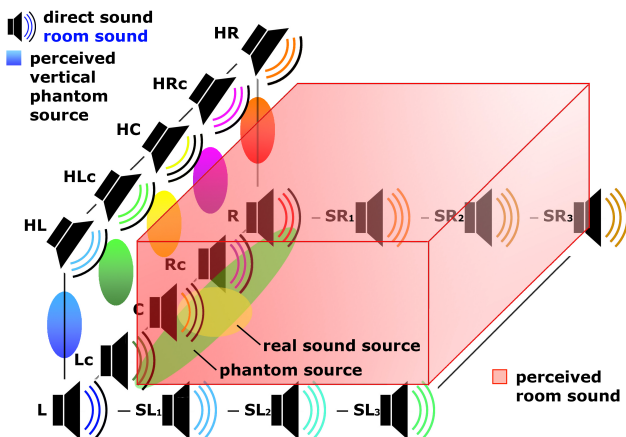


Fig. 14: In this extended 9.1 loudspeaker setup, five stereo and equidistantly arranged stereo pairs in the front represent instruments / sounds. All loudspeakers reproduce room sound and thus lead to an impressive envelopment.

If loudspeakers are placed at head height in the reproduction room, the listeners closest to the loudspeakers cover the sound reproduced by the loudspeakers, in particular of the high frequencies. As a result, only the front listeners can hear the front lower loudspeakers, and only the rear listeners can hear the rear lower loudspeakers unlimited. Thus, an imbalance occurs for most listeners. It is therefore more appropriate to arrange the lower loudspeakers slightly above the heads of the listeners to avoid such shading effects.

## 3. Applied psychoacoustics in the signal assignment for a 3D audio PA system

The previous chapters explain the optimal design for an economical 3D audio loudspeaker layout and basic considerations for the required loudspeaker signals. The next step is to consider how sound engineers have to assign microphone signals, samples, etc. to the many loudspeakers from a psychoacoustic point of view.

### 3.1. Signals for the envelopment

#### 3.1.1. Native captured room sound

In many cases, captured room sound is the best choice for a plastic-sounding envelopment. The following conditions exist for this:

- The recording room must be acoustically excellent. Room sound from small rooms or rooms with acoustic problems are much more noticeable if reproduced by a 3D audio system than with mono and stereo. This means that such signals lead to bad-sounding recordings.

- The room signals must not have direct sound, especially at high frequencies. Otherwise, especially in the back of the listening area, listeners perceive sound sources from the direction of the direct sound components. In many cases, this leads to double and multiple images or increased muddiness.

#### 3.1.2. Artificially generated room sound

It is not easy to provide room microphone signals with correlation around zero in the required number, especially for larger loudspeaker setups. Room sound processors and upmixing tools represent an alternative. As with the room microphone signals, important conditions must also be considered:

- Simple reverb units and plug-ins have a comparatively low quality for confounding real room sound. Only the best reverberation processors generate satisfactorily appearing room sound [8].
- Upmixing tools have the challenge of extracting spatial components from every possible existing sound material. Therefore, they are limited in the generation of plastic sounding room sound.

#### 3.1.3. Combination of native room sound and upmixing

Especially in the case of already realized recordings, there are in many cases fewer captured room signals than are needed for sound addressing in 3D. One solution is the combination of existing room sound signals with an upmix of the original signal, compare [8].

#### 3.1.4. Pad and noise-like sounds

In synthetically produced music there is usually no room sound. In most cases, room sound is also out of the question for that kind of music. Alternatively, pad and noise-like sounds can be used for the envelopment [6]. In many cases, it is sufficient to use multiple instances of the same sound to generate signals in sufficiently high numbers of channels with correlation around zero [8].

### 3.2. Assigning sound sources to loudspeakers

In this article, the author limits himself to the case of direct sound from the front. Direct sound from other directions as well as panning direct sound is not discussed here.

#### 3.2.1. Mono sound sources

As described in chapter 2.5.1., mono sound sources are especially suitable for use as real sound sources. There are two cases:

- Cocktail party effect: When two or more real sound sources are present at different locations in the room, they are clearly distinguishable from each other and do not cover each other or only weakly [6].

- Sense of depth: If two or more mono sound sources appear at the same point in space, masking and merging effects occur [6]. Depending on the level ratios and spatial proportion of individual sounds, a pronounced sense of depth occurs [9].

### 3.2.2. Stereo sound sources

Most acoustic instruments and in particular string instruments radiate direct sound with different spectra depending on the direction. In addition, resonant bodies represent their own room, which behaves acoustically similar to a recording room [10]. Stereo close-miked instruments of this type therefore always have partly direct sound, some room sound, which cannot be separated easily. The stereo information of the signal is retained if the assignment to the loudspeakers is two-channelled. There are basically two possibilities for an assignment:

- Horizontal phantom source: At this assignment, the imaged sound source acts as an audible connection between instruments due to the expansion of horizontal phantom sources [8]. At the same time, masking and merging effects occur, allowing sense of depth. As described in chapter 2.5.1., different widths of stereo bases can be used during the mixing process.
- Vertical phantom source: At this assignment, the imaged sound source is spatially delimited from those with other directions. As a result, on the one hand, the transparency is very high. On the other hand, vertically stereo reproduced instruments sound more natural than horizontal phantom sources [8].

## 4. Conclusion

The cost of a live event PA system with 3D audio is very high compared to a 9.1 loudspeaker setup due to the even higher number of required loudspeakers and playback channels. By applying psychoacoustic phenomena, however, it can be kept to a reasonable level.

The spatial resolution for a natural sounding image is phenomenal, especially at larger loudspeaker setups with many front loudspeakers or vertical stereo pairs. So there is not just the chance of obtaining a large part of the psychoacoustically occurring phenomena in an upscaling, but due to the much higher resolution of sound sources in the front, the reproduction of music using 3D audio is also much closer to natural hearing with a maximum localization sharpness of about 1°!

## 5. References

- [1] Nipkow, L.: Properties of Auro 3D room signals. 27. Tonmeistertagung, Cologne (2012), Proceedings, ISBN 978-3-9812830-3-7; 757-770
- [2] Blauert, J.: Räumliches Hören. S. Hirzel Verlag, Stuttgart, 1974, ISBN 3-7776-0250-7
- [3] Blauert, J.: Räumliches Hören. Nachschrift: neue Ergebnisse und Trends seit 1972. S. Hirzel Verlag, Stuttgart, 1985, ISBN 3-7776-0410-0
- [4] Lee, H.: 'Sound Source and Loudspeaker Base Angle Dependency of the Phantom Image Elevation Effect' Journal of the Audio Engineering Society, 65 (9), pp. 733-748 (2017). ISSN 1549-4950
- [5] Lee, H.: 'Phantom Image Elevation Explained'. In: Audio Engineering Society the 141st International Convention, Los Angeles, USA (2016)
- [6] Nipkow, L.: Sound Design in 3D for Dance Music. 4. ICSA, Graz (2017), Proceedings
- [7] Potratz, U.: Untersuchung der Gestaltungsmöglichkeiten früher Reflexionen mit Hilfe eines raumakustischen Modells. Diplomarbeit, Erich-Thienhaus-Institut der Hochschule für Musik Detmold (2005), <http://www.eti.hfm-detmold.de/lehraktiv/diplomarbeiten/diplomarbeitenorder/da-potratz.pdf>
- [8] Nipkow, L.: The Importance of 3D in Immersive Audio. 30. Tonmeistertagung, Cologne (2018), Proceedings, ISBN 978-3-9812830-9-9
- [9] Nipkow, L.: Sense of depth in 3D for Pop Music. 29. Tonmeistertagung, Cologne (2016), Proceedings
- [10] Nipkow, L.: Room signals – properties and influence on Auro 3D recordings. 2. ICSA, Erlangen (2014), Proceedings, ISBN 978-3-98 12830-4-4, 96-101



# Spatial audio production for immersive fulldome projections

Presented \* by VDT.

Johannes Ott<sup>1</sup>, Anca-Stefania Tutescu<sup>1</sup>, Niklas Wienböcker<sup>1</sup>, Jan Rosenbauer<sup>1</sup>, Thomas Görne<sup>1</sup>

<sup>1</sup> *Hamburg University of Applied Sciences (HAW Hamburg)*

*Department of Media Technology*

*Email: johannes.ott@haw-hamburg.de*

## Abstract

Full dome is an immersive half-spherical video format utilized mainly in planetariums which is often combined with spatial audio playback. This combination on one hand offers new ways of perceiving sound in space and on the other hand helps enhancing full dome productions with audiovisual synergy. However, audio production for full dome video poses some technical and artistic challenges. Limited time slots and resources seldom allow to work on sound productions inside a planetarium directly. Likewise, the various spatial audio technologies provide the user with fairly different approaches to create, position and move sounds in space.

This paper investigates three different approaches to create spatial audio content for full dome productions in remote studios and to present them in a planetarium: object based proprietary Fraunhofer “SpatialSound Wave” (SSW) system, scene based Higher Order Ambisonics (HOA), and channel based production will be compared. Technical challenges and potentials of storytelling in spatial audio will be discussed.

## 1. Introduction

The terms “fulldome” or “fullspace” refer to half spherical video projections designed to present the audience with an immersive multi-media experience in venues like planetariums, resembling illusionistic Renaissance ceiling paintings (Gerling, 2013). Today fulldome productions form an increasing part of multi-media shows in planetariums worldwide. In the wake of this new immersive content, planetariums have started to recognize the potential of spatial audio and have installed suitable playback systems in their domes.

Although the workflow for producing spatial audio content for fulldome is very similar to other immersive formats such as fixed media concerts, sound installations or VR, playing it back in a planetarium brings up a variety of challenges:

- The size of the room is very likely much larger than the listening room in which the content was produced. Even

though their size is similar to cinemas, no guidelines or standards exist for audio playback in planetariums, which makes it difficult to judge sound characteristics (such as loudness) in the production stage.

- Because of the floor area occupied by the star projector – the essential piece of technology in a planetarium that is always installed in the very center of the dome – seats are usually grouped around the center, leaving open the area where the ideal listening positions would be.
- To maximize seating capacity, seats in planetariums generally go right up to the walls of the room, giving the audience seated there necessarily a different auditory experience (see Fig. 2, p. 5).

Dealing with these challenges, this paper investigates workflows when producing with generically different spatial audio representations (object based, scene based, channel based), discussing their assets and drawbacks.

## 2. Production formats

### 2.1. Object based production (SSW)

The proprietary “SpatialSound Wave” (SSW) System developed at Fraunhofer IDMT has been installed in several planetariums in Germany, such as Hamburg, Jena, Berlin and Bochum (Fraunhofer IDMT, [n. d.]). Producing in this format is therefore an obvious choice for fulldome production.

SSW is an object based / hybrid system based on the works of Brandenburg et al. (2013) with its origin in Wave Field Synthesis (WFS). The core of the System is a real-time renderer that spatializes monophonic sound sources for multi-channel loudspeaker arrays. The audio content is fed into the renderer from an external playback machine.

Spatialization in SSW is controlled wirelessly via a web interface in real time. To produce SSW content, a workstation for playback of the mono sound sources, the SSW renderer and a loudspeaker array are needed. Playback through headphones is not possible. The movement of sound sources is synchronized with the playback machine via timecode.

SSW productions can be delivered to other SSW systems in the form of 32 mono tracks containing the audio information of the sound sources and the SSW session containing the positional meta-data. If the playback venue has a SSW system installed the studio production translates easily to the playback system. To compensate for the difference in size of studio and venue, a scaling factor can be applied. It is also possible to pre-render the loudspeaker signals for a specific playback system by recording the outputs of the rendering unit.

### 2.2. Scene based production (HOA)

A convenient way of producing sound for a multi-directional medium like fulldome is to use auditory scenes, in this case achieved by utilizing Higher Order Ambisonics (HOA). This technology, based on the works of Gerzon (1973), has become very popular in the spatial audio scene, not least due to format specific advantages of scene based coding like e.g. the possibility of rotating the soundfield. Lots of solutions exist today for producing Ambisonics content and many of them are open-source. For linear content such as fulldome, working with tools like VST plugins in a DAW differs the least from conventional linear audio production, and the resulting signal takes full advantage of the spatial resolution of the playback venue given that the utilized Ambisonics order is high enough.

Further advantages of producing in HOA are that no specific hardware except for the loudspeaker array is needed, and that the signal can be rendered for virtually any playback situation including stereo or surround, i.e. the Ambisonics format can be regarded as system-agnostic. This is helpful since the number of planetariums equipped with spatial audio playback systems is still small. The content can also easily be adapted for binaural playback and other applications. It is theoretically even possible to produce on headphones, which reduces the technical demands for producing for planetarium to a minimum compared to the other approaches discussed.

#### Ambisonics and HOA

Higher Order Ambisonics (HOA) should not be confused with first order Ambisonics, created e.g. with the now popular tetrahedral microphones. First order Ambisonics is known for suffering greatly from sweet spot issues, as the spatial auditory scene tends to collapse when the listener is situated at a non-ideal position. We therefore would not recommend it as a format for fulldome productions. In contrast, HOA productions – specifically from 5th order upwards – are rather robust regarding the listening position, as shown by Frank and Zotter (2017). The authors’ experiences conform with these findings.

Ideally the playback facility would be equipped with an Ambisonics renderer configured for the playback system. This renderer could easily be implemented with an ordinary computer powerful enough for HOA rendering, running e.g. a DAW software with a dedicated plugin like the AllRADecoder of the IEM Ambisonics plugin suite (Zotter and Frank, 2012). Setting up a HOA renderer like this would be a low-cost yet very valuable addition to any planetarium’s technical setup.

In case the HOA signal cannot be decoded in real time, i.e. if no Ambisonics renderer is available in the venue, it is also possible to pre-render for the specific playback system of the venue, creating a multi-channel loudspeaker feed. In this case the loudspeaker signals would be played back directly by short-cutting a potentially installed 3D audio processor like SSW. However, this direct access to the loudspeakers might not be available in every venue.

In case these two “best options” are not realizable, a less optimal workaround is possible, assuming the venue is equipped with SSW: the HOA signal can be decoded to a virtual loudspeaker array created within the SSW system. The obvious way would be recreating the production venue’s setup as a virtual loudspeaker dome in SSW. The alternative way would be to render the HOA in a last step of the production process to a virtual setup matching the loudspeakers in the playback venue, as if it would be rendered for direct HOA playback. Unfortunately, either way requires two layers of virtualization which can result in a significant degradation of localization compared to other workflows. Although the latter solution might appear needlessly complicated (as virtual point sources would be positioned at actual loudspeaker positions), this might be worth trying for the reward of a less colored and more stable soundfield.

A factor to be taken into account for scene based audio is the listening position, which can vary greatly depending on the size of the venue and space occupied by projectors or similar. While in theory the sweet spot – the “perfect reconstruction area” – for Ambisonics is very small, listener feedback suggests a satisfying listening experience and stable scene perception in an extended radius outside the sweet spot, specifically when working with Higher Order Ambisonics (Frank and Zotter, 2017).

### 2.3. Channel based production

The production workflow with a channel based approach differs significantly from those discussed previously. It is assumed that the production will be played back on a known loudspeaker array. Sounds might be distributed over several speakers using e.g. Vector Base Amplitude Panning (VBAP) after Pulkki (1997), or they might be static using dedicated loudspeakers for the different sonic objects (which in fact means utilizing the playback system as a “loudspeaker orchestra” in the tradition of spatialized electro-acoustic music; cf. e.g. Brech and Paland 2015; Voit 2014). Either way such a production benefits greatly from a High Density Loudspeaker Array (HDLA) with 20+ loudspeakers.

Advantages of a HDLA channel based approach are the very precise localization of sounds with clearly localizable auditory objects, virtually no spatial blur, no sweet spot restriction (if no panning is applied), and perfect control of spatialization assuming there is a sufficient number of loudspeakers available.

Of course, in less avant-garde environments, “channel based production” usually refers to material produced in and for a well-defined comparatively sparse array, like Auro-3D 9.1. The latter or similar 3D Audio systems can be classified as “surround with height”, with one sole height layer above ear level, and with the height loudspeakers typically situated vertically above the ear-level loudspeakers, which complicates triangular panning like VBAP and leads to unstable phantom sources / auditory objects at the sides, rear and height (for a list of channel based systems see ITU-R BS.2051 2018). The advantage of a 3D 9.1 or similar approach would be at least the availability of production facilities.

However, since there is no standardized loudspeaker array for planetariums, playback for channel based fulldome productions is difficult. Ideally a channel based production would mean to produce directly at the location, i.e. in the planetarium dome. This is very unlikely because of time constraints of a running planetarium, and furthermore this would lead to a production playable just in one particular planetarium. So it comes back to the task of matching a channel based studio production – be it common 3D 9.1 or fancy 33.2 – to a different loudspeaker array.

A rather pragmatical way of dealing with non-matching loudspeaker setups would be choosing the “nearest to perfect” loudspeakers from the array. Here the task would be – similar to playing back decoded HOA – to get direct access to the loudspeakers. And the significance or insignificance of spatial errors and coloration introduced by the utilization of a non-ideal setup are hard to predict.

A practical approach to adapt a channel based spatial production to different playback systems is to render the channels within the respective playback system (like SSW or HOA) as virtual loudspeakers. In theory this would preserve the spatial and sonic properties of the production, but in practice some advantages of channel based production might get lost as spatial blur and coloration of the virtual loudspeaker system are introduced.

### 3. Storytelling and formal considerations in spatial audio

A common assumption in content production for so called “immersive media” is that they facilitate a deeper experience of immersion and stronger emotional impact compared to conventional audiovisual media (see e.g. Hahn 2018; Uhrig 2015). While it is widely believed that the immersive experience provided by media like fulldome, 360° video or VR might be explained by an enhanced “realism”, it should be noted that this technically mediated realism – specifically with fictional content – is basically different from the real-world experience, as the audience is always aware of its fictional nature however strong its emotional impact might be; Carroll coined the term *paradox of fiction* (Carroll, 1990; Voss, 2009).

Following Ijsselstein et al. (2000), *presence* – a more specific description of the immersive experience – can be understood as feeling present in a mediated environment and losing awareness of the technological medium involved. Thus an objective of media production for fulldome would be enhancing the audience’s immersion and experience of presence in a virtual space – realistic or not – by technical means.

But what creates a believable and effective virtual auditory environment, both in fulldome productions and in other media such as VR and cinema? Lennox et al. (2001) argue that realism is based not solely on the directionality of sound and geometrical representation of space, but rather on the relationship between sound objects and their environment, i.e. their sonic context.

Spatial audio systems, for their advanced options to arrange and direct auditory objects in space, bring added spatiality to storytelling, being a powerful tool for better directing the audience’s attention and increasing presence, according to Hendrix and Barfield (1996).

Barrett (2010) states that a technically sophisticated spatial audio system – her example being Higher Order Ambisonics – facilitate realism beyond sheer listener envelopment and might be capable of even “*breaking the paradox between reality and fiction*”.

And Ijsselstein et al. (2000) note spatial sound as a conveyor of sensory information that increases presence depending on fidelity and extent of auditory events. These sound properties do not necessarily mean real-world accuracy.

According to Grimshaw (2014), immersion is further improved via referencing other senses (touch, smell, taste) through sound: “*A visually rendered corpse alone may not trigger an olfactory sensation of decay, but the sound of buzzing flies and wriggling maggots around it has a much greater potential to do so. [...] Although a realistic virtual soundscape may (virtually) reflect its reality counterpart, the lack of audio input compensating for other sensory modalities creates an incomplete experience, lacking immersion and, ironically, appearing unrealistic.*” (Grimshaw 2014: 368)

Lennox et al. (2001) talk about creating a sound hierarchy: “my space”, “adjacent space” and “distant space”, with the



first containing the closest and most urgency-filled sounds, the second being less relevant in terms of urgency or threat to the listener, but having the potential to move into their close space, and distant space providing the least amount of threat and requiring the least amount of localization to be believable (cf. Hall's well-established concept of social distances, see Hall 1969). Context is hereby a requirement for the ability to employ selective attention to distinct objects and features, and the authors speak of "selective inattention" with respect to background information in the environment, required in order to detect consistency or inconsistency in space. Likewise, Dalton and Fraenkel (2012) investigate "inattentive deafness", describing the effect of not consciously perceiving a sonic event if the attention is captured by a different part of the auditory scene. Nonetheless, one can expect that the "overheard" elements of the auditory scene have an emotional impact as well (Bargh, 1988).

Movement is also an important source of information and can be expressed not only by change in location, but also through spectral changes like high-frequency comb filtering (Lennox et al., 2001). Consequently, movement should be considered as a key element of a compelling spatial audio production (cf. Karadoğan and Görne 2019).

Spatial sound productions in planetariums allow for novel approaches to storytelling. Cinematography for such a medium brings challenges such as the audience's orientation in space or the risk of disorienting and making the viewers sick through jump-cuts and sudden panning because of the fact that the viewers' field of view is enveloped, unlike with traditional screens (Yu et al., 2007).

The role sound plays and its relationship to the visuals are also in question. One of the tasks sound can fulfill is directing the audience's attention via auditory cues. Although this is a common stylistic device in film sound design (Görne, 2017) it is crucial for sound design in full-dome video (and comparable to other immersive media like VR or 360° video) because of the visual envelopment, even though the viewers are rather stationary in this case, encouraged to lean back their chairs and being set up in a way that creates roughly the same preferred spot of viewing (even if from different directions).

An experiment done by Sheikh et al. (2016) employed 360° video and stereo sound to divert attention from the main characters to a third person in the scene, employing a variety of means (gestures from characters, the main characters addressing the viewer directly, another person walking to the target): *"Audio cues have the advantage that no assumption is made about the viewer's focus of attention at the time of the cue. Even without fully spatialised audio, the use of sound also alerts the viewer that there is something to see; with the visual cue alone, participants sometimes followed the cue, but not as far as the target. When both audio and visual cues were used, all participants saw the target."* (Sheikh et al., 2016)

Sound can be tied to the picture in a straightforward (hyper-)realistic fashion, or it can act independent from it, which French (2018) claims to increase extended presence in full-dome productions through viewer participation and exchange

with the medium and its different sound layers and objects (for an extended discussion on sound/image relations in film see Görne 2017).

An example of an approach that tries to go one step further, taking advantage of the capabilities of a planetarium, is an audio-visual piece produced by some of the authors in 2018. It was attempted to employ a unique storytelling approach: half of the production was meant to be experienced with the eyes closed, to communicate events from the story via flashes of light and color and basic geometric shapes that simulate what we normally perceive visually in that state. As this approach has its limitations due to the default brightness of the video projection when displaying pure black (which makes it hard to simulate directional light and shapes), spatial sound was the principal medium conveying the story, but the light did have an interesting function in maintaining a visual connection to the medium. The second half of the piece was realized as a conventional full-dome video with spatial sound.

The concept was to put the audience in the shoes of someone "blindly" going through his or her last day before suffering a car accident, at which point an experience of the "afterlife" or an altered reality, reassembled from bits and pieces of the last day's memories, was introduced. In between, the "transitioning" was marked by a star projection using the planetarium's Carl Zeiss Universarium Projector.

## 4. Case studies

In 2018 and 2019, two evening programs under the title "Equinox" were produced at HAW Hamburg's Immersive Audio Lab (IAL) for screening at Planetarium Hamburg. They consisted of several distinct full-dome productions with spatial audio. Most tracks were produced in 7th order Ambisonics while two were produced using a channel based approach.

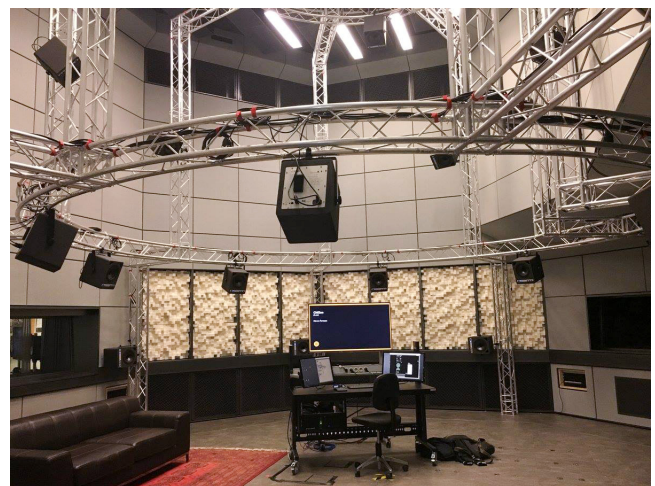


Fig. 1: Situation at IAL (without ring 0 / floor loudspeakers).

The production system at IAL is an 33.2 array arranged in 5 height layers with ring 1 being at ear level (Fig. 1, Table 1). The array layout was designed to accommodate not just HOA beyond 6th order in half-space but also a wide range of channel and object based audio coding formats

Ring	Height	Diameter	Channels
0	-1,0 m	6,5 m	4
1	0 m	8,0 m	8
2	+1.2 m	6,5 m	8
3	+2.4 m	6,5 m	8
4	+3.5 m	4 m	4 + 1 (VoG)

Tab. 1: IAL array layout, height relative to ear level.

including setups commonly used in sound art and electro-acoustic music; for details see Kessling and Görne (2018).

The playback system at Planetarium Hamburg is a 60.4 array arranged in 4 height layers plus one layer solely for the “Voice of God” loudspeaker, with a dense array in the horizon ring (nominally ear level) and rather sparse in the height layers (Fig. 2, Table 2).



Fig. 2: Planetarium Hamburg. Note the dense horizon array (ring 1).

Ring	Height	Diameter	Channels
1	0 m	20.6 m	36
2	+4.0 m	19 m	11
3	+7.3 m	14.5 m	8
4	+9.5 m	7.9 m	4
VoG	+10.3 m	-	1

Tab. 2: Planetarium Hamburg array layout (level 1 at nominal height of 0 m equals the synthesized ear level by the SSW system; actual height is some 1...2 m above the listeners’ heads).

In addition to the screenings at Planetarium Hamburg, the productions were also submitted to the 12th and 13th Fulldome Festival Jena Student Competition. Therefore they had to be prepared for two different playback systems.

In 2016 one of the authors produced a fulldome video utilizing a SpatialSound Wave system at Hochschule Darmstadt’s Soundscape & Environmental Media Lab (SEM). The SEM-Lab is equipped with a flexible and mobile loudspeaker array which at the time consisted of 24.2 loudspeakers. This production was shown at the 10th Fulldome Festival Jena Student Awards. It was also adapted in 5.1 for screening at a mobile fulldome tent at Hessentag 2017.

## 4.1. SSW content

This is the most straightforward approach, assuming the planetarium has SSW installed.

The external spatialization of sounds makes the system beginner-friendly. In contrast to technologies like Ambisonics, no deeper understanding of underlining principles is needed to set up a session for spatialization. The user can connect her or his mobile device to the renderer and control all sound sources at once, which makes it very easy to spatialize pre-produced mono tracks – representing the sound objects – from inside the planetarium.

While this approach is very versatile for live applications, it has disadvantages in linear productions such as fulldome video. Movement can only be recorded in real-time, and the automation of movement cannot be edited as it is common in DAW software. This makes operations such as deletion, copy-pasting and adjustments time consuming and imprecise. Time constraints are therefore a big factor if the producer is spatializing her or his pre-rendered sound sources in the planetarium. If the studio used for the production is not equipped with its own SSW renderer, some of the other workflows discussed might be advisable.

The playback of SSW productions in other venues with SSW systems is relatively easy. However different room sizes might make sound adjustments necessary. Since they are transmitted as separate mono channels, each of the different sound objects can still be adjusted in volume and with effects like e.g. equalization, which is not the case for other approaches such as Ambisonics. The transfer in the form of multiple mono files is also susceptible to mistakes by both the producer and the planetarium technicians. It is crucial that a shared naming convention for all files should be discussed beforehand.

If the production is to be shown in a venue without a SSW system it is possible to pre-render the audio channels for the respective playback system. This is achieved by configuring the SSW renderer in the studio with the speaker positions of the venue and recording its outputs. However, this requires changing the routing of the studio and knowledge of the exact speaker positions of the venue. It is also a real time process. Changes during sound check are not possible anymore.

While rendering for other formats such as 5.1 is theoretically possible, in the case of our production playback on this much smaller horizontal array could not achieve a good representation of localization and sound. To preserve the artistic intent, a dedicated manual remix/downmix for this format would be preferred. Further experiments with productions in Ambisonics would be of interest.

## 4.2. Higher Order Ambisonics (HOA) content

### 4.2.1. Producing and playing back in HOA

This is the second most straightforward approach, as producing with an open system like HOA is very convenient.

HOA productions were made using Cockos Reaper DAW and the Ambisonics plug-in suite developed by Zotter et al. and

Rudrich at IEM (Institute of Electronic Music and Acoustics) Graz (Frank et al., 2015; Rudrich, 2018; Zotter and Frank, 2012). The IEM StereoEncoder was our primary plug-in to pan channels and write automation for spatial movement. Productions were made in 7th order to take advantage of the full spatial resolution of the IAL loudspeaker dome.

For the presentation in March 2019 we configured IEM's AllRADecoder with the planetarium's speaker positions, rendering the signals for the loudspeaker array in Hamburg planetarium, and connected our playback laptop via Dante, short-cutting the renderer in the venue and directly feeding the pre-produced loudspeaker signals to the playback system. This experiment basically worked quite well, as can be expected with a system-agnostic format like HOA.

However, the perceived locations of auditory objects might be displaced drastically due to precedence effect if single loudspeakers are too close to the audience – a factor we found to be of special importance in planetariums: The seats close to the “rear” wall of a planetarium with the lowest loudspeakers very close above can be considered the worst seats in the venue, where it might happen that the audience perceives everything behind and above the head due to precedence effect. Of the approaches discussed here Ambisonics can be expected to suffer the most from this effect, since all the loudspeakers are contributing equally to the sound field even if just a single point source is rendered.

Disregarding speakers that are too low and too close, thus utilizing just a smaller slice of the system, increases the distance to the nearest speakers for these seats and thus possibly prevents the precedence effect at these positions. Even though the utilization of less speakers might result in a lower effective Ambisonics order and therewith a smaller listening area and lower spatial resolution, we found that trading off Ambisonics order against precedence effect by practically switching off loudspeakers helps to even out the listening experience for all seats of a planetarium. Experimenting with the lowest loudspeakers (i.e. the horizon ring) combined with careful listening at different seats in the venue is greatly encouraged. An alternative might be to keep the seats close to the room boundaries unoccupied.

#### 4.2.2. Adapting HOA to SSW

As easy as it is to render HOA content for the loudspeaker array of the playback venue, as complicated it gets when HOA has to be played back via SSW. This might be necessary if the venue's signal routing cannot be restructured and direct loudspeaker access is not available before the show. We realized the HOA/SSW adaptation in March 2018 at Planetarium Hamburg and for both screenings at Planetarium Jena. Basically two workflows can be implemented:

1. decoding HOA to the loudspeaker array in the planetarium (which then would be substituted with fixed SSW objects),
2. creating SSW sound objects representing the loudspeaker positions of the production venue,

Both options utilize SSW as a static playback system, where

movement of auditory objects is implemented in the HOA signal. The first option is similar to rendering directly for the loudspeakers in the planetarium, with an extra SSW processing stage in between, which seemed to be rather inelegant. We therefore opted for the second approach, creating a virtual IAL loudspeaker dome from SSW objects in the planetarium. It turned out that the “bottleneck” of multiple format conversions (sound object → HOA encoding → HOA decoding → SSW encoding → SSW decoding) led to a substantial spatial imbalance, so that the content had to be remixed for the final presentations with a specific focus on loudness of auditory objects above the horizon ring. Another problem is the decreased number of virtual loudspeakers available. Even though SSW can synthesize 32 sound-objects at once, the setup in both planetariums allowed to use only 24 of them. This was a significant degradation of spatial resolution from the 33 loudspeaker array at IAL.

### 4.3. Channel based content

Here, “channel based production” refers to pre-produced material that has been rendered within a specific loudspeaker array like Auro-3D or similar, with one exception described below. Direct playback of such material is possible only if a subset of the loudspeaker array in the planetarium matches the original production array, at least approximately (which for example is not working for Auro 9.1 playback in Hamburg planetarium).

#### 4.3.1. Live spatialization

A workflow similar to the customs of sonic art and electro-acoustic music is the live spatialization of more or less pre-produced tracks, the presentation being a live performance and the loudspeaker array used as a “spatial diffusion system” or “loudspeaker orchestra”. If the venue has a system like SSW with its capabilities of live processing installed, then the pre-production would typically be channel based, where during the live event the channels are treated as sonic objects.

#### 4.3.2. Adapting channel based material to SSW

This adaptation is similar to our approach of adapting HOA to SSW, but with a significantly lower number of virtual loudspeakers (e.g. 9 or 11 for Auro-3D content) implemented as SSW objects. Hence the spatial impression is less stable, and possible aberrations of the virtual loudspeakers tend to have a larger impact.

#### 4.3.3. Adapting channel based material to HOA

Adapting a standard 3D Audio production to HOA is similar to adapting to SSW. However, in our experiments with productions for Hamburg planetarium, we also dealt with channel based productions in our in-house 33.2 format, which then were rendered as 7th order HOA to achieve system compatibility. Results were quite satisfactory; there was no obvious difference between generic channel based and generic HOA material when rendered as HOA for the planetarium loudspeaker array.

## 5. Conclusion

Full-dome productions benefit greatly from spatial audio. Our experiments with dedicated spatial audio compositions and sound designs for full-dome video have been received very positively by the audience, independent of production and playback formats. The extended storytelling options and enhanced immersive experience of spatial audio for full-dome by far outweigh practical issues and more or less complex workflows.

The Fraunhofer SpatialSound Wave system has become a de facto standard in this niche. If a studio equipped with SSW is available, producing in this format makes playback in planetariums with SSW straightforward. However, this requires spatialization outside the DAW with its convenient editing features, and limits the distribution to venues with SSW systems.

For playing back channel based content in planetariums there are several options; however, all of them require some preparation. Channel based material with a rather low spatial resolution and a low number of channels like 9.1 might be presented as “surround with virtual height”, but due to the few virtual height loudspeakers issues with sound coloration and with spatial resolution can be expected. Of course, such a “3D” format would still be an advantage over the common stereo or surround playback in the venue, yet immersion, presence and emotional impact might not be as impressive as one would expect from spatial audio.

Producing in HOA provides the producer with a rather familiar workflow similar to conventional audio productions when compared to SSW. Furthermore, HOA leaves all options of distribution open, as the playback venue does not need to have an Ambisonics system installed. Preparing a HOA production for playback in a remote venue is uncomplicated: in case the venue is not equipped with a HOA rendering system, the producer simply needs the geometrical data of the playback array to render loudspeaker signals for the specific venue (and then the only technical obstacle to overcome would be finding a way to feed signals directly to the loudspeakers).

If the production was made in HOA then of course HOA playback is advisable, as the adaptation to a system like SSW indeed is possible, but likely leads to degradation of the content.

A yet still pending experiment for the production of HOA material is the virtual spatialization in the studio utilizing binaural rendering in HOA instead of the loudspeaker dome. A production environment like this could make full-dome productions realizable even for smaller studios. We expect promising results, especially when personalized HRTFs are incorporated.

The choice of technology thus depends on the requirements of the full-dome production itself: on the venues where it will be shown (only one particular planetarium, only planetariums with SSW, other loudspeaker arrays outside of the full-dome niche), and finally if it will be edited for other distribution channels such as 360° video or VR.

One last remark on the production of spatial audio content: in conventional music, audio drama or sound design productions it is very common to think of sound production and “spatialization” (typically in form of a stereo or surround mix) as independent production steps. However, from our experience this does not apply to spatial audio, as the spatialization here becomes a substantial part of the artistic process (see e.g. Karadoğan and Görne 2019). There is a remarkable difference between a workflow where the sounds are produced in a conventional studio and then “mixed” in the loudspeaker dome, and a workflow where the composition is made in and for the immersive environment.

## Acknowledgement

Most full-dome video works described in this paper have been produced during winter semester 2017/18 and 2018/19 in the classes of Prof. Almut Schneider (experimental video), Prof. Mareike Ottrand (interactive illustration) and Gloria Schulz (video technology / video mapping) of Hamburg University of Applied Sciences.

The described SpatialSound Wave production for full-dome video was produced during the winter semester 2015/16 under the supervision of Prof. Sabine Breitsameter of Darmstadt University of Applied Sciences.

A special thanks goes to Prof. Thomas W. Kraupe, Principal and Artistic Director of Planetarium Hamburg, and Sascha Kriegel, Technical Director of Planetarium Hamburg, for extensive support and for not just welcoming but stimulating technical and artistic experiments in their house.

Photo credit: Gertje König, Hamburg.

## References

- John A. Bargh. 1988. Automatic Information Processing: Implications for Communication and Affect. In *Communication, Social Cognition, and Affect*, Lewis Donohew, Howard E. Sypher, and E. Tory Higgins (Eds.). Lawrence Erlbaum Associates.
- Natasha Barrett. 2010. Ambisonics and acousmatic space: a composer’s framework for investigating spatial ontology. In *Proc. 6th Electroacoustic Music Studies Network Conference*. Shanghai.
- Karlheinz Brandenburg, Martin Schneider, Andreas Franck, Walter Kellermann, and Sandra Brix. 2013. Intelligent Multichannel Signal Processing for Future Audio Reproduction Systems. Conference Paper, AES 52nd Int. Conference on Sound Field Control – Engineering and Perception, Guildford.
- Martha Brech and Ralph Paland (Eds.). 2015. *Kompositionen für hörbaren Raum – Die frühe elektroakustische Musik und ihre Kontexte*. transcript.
- Noël Carroll. 1990. *The Philosophy of Horror. Or: Paradoxes of the Heart*. Routledge.

- Polly Dalton and Nick Fraenkel. 2012. Gorillas we have missed: Sustained inattentive deafness for dynamic events. *Cognition* 124, 3 (2012).
- Matthias Frank and Franz Zotter. 2017. Exploring the perceptual sweet area in Ambisonics. Convention Paper, AES 142nd Convention, Berlin.
- Matthias Frank, Franz Zotter, and Alois Sontacchi. 2015. Producing 3D Audio in Ambisonics. AES 57th Int. Conference, Hollywood.
- Fraunhofer IDMT. [n. d.]. IDMT SpatialSound Wave Website. [www.idmt.fraunhofer.de/de/institute/projects-products/spatialsound-wave.html](http://www.idmt.fraunhofer.de/de/institute/projects-products/spatialsound-wave.html). ([n. d.]). Accessed: 2019-08-29.
- Michaela French. 2018. Using the layers of presence as a framework for artistic practice in fulldome space. In *IPS 2018 Toulouse Conference Proceedings*.
- Winfried Gerling. 2013. Die Kuppel als medialer Raum. In *Fullspace-Projektion. Mit dem 360° Lab zum Holodeck*, Gordian Overschmidt and Ute B. Schröder (Eds.). Springer Vieweg.
- Michael A. Gerzon. 1973. Periphony: With-Height Sound Reproduction. *Journal of the Audio Engineering Society* 21, 1 (1973), 2–10.
- Thomas Görne. 2017. *Sounddesign. Klang, Wahrnehmung, Emotion*. Hanser.
- Mark Grimshaw (Ed.). 2014. *The Oxford Handbook of Virtuality*. Oxford University Press.
- Ephraim Hahn. 2018. Musical Emotions Evoked by 3D Audio. Conference Paper, AES Conference on Spatial Reproduction – Aesthetics and Science, Tokyo.
- Edward T. Hall. 1969. *The Hidden Dimension*. Anchor Books.
- Claudia Hendrix and Woodrow Barfield. 1996. The sense of presence within auditory virtual environments. *Presence: Teleoperators and Virtual Environments* 5, 3 (1996), 290–301. <https://doi.org/10.1162/pres.1996.5.3.290>
- Wijnand A. Ijsselstein, Huib de Ridder, Jonathan Freeman, and Steve E. Avons. 2000. Presence: Concept, determinants and measurement. In *Proceedings of SPIE - The International Society for Optical Engineering*, Vol. 3959. <https://doi.org/10.1117/12.387188>
- ITU-R BS.2051. 2018. Advanced sound system for programme production. Intern. Telecom. Union Radiocom. Sector Recomm. ITU-R BS.2051-2. (2018).
- Can Karadoğan and Thomas Görne. 2019. Auditory Scenography in Music Production: Case Study Mixing Classical Turkish Music in 7th Order Ambisonics. Conference paper, AES Conference on Immersive and Interactive Audio, York.
- Philipp Kessling and Thomas Görne. 2018. Studio for immersive media research and production: Immersive Audio Lab at HAW Hamburg. e-Brief EB453, AES 145th Convention, New York.
- Peter P. Lennox, John M. Vaughan, and Tony Myatt. 2001. 3D Audio as an Information-Environment: Manipulating Perceptual Significance for Differentiation and Pre-Selection. In *Proceedings of the 2001 International Conference on Auditory Display, Espoo*.
- Ville Pulkki. 1997. Virtual Sound Source Positioning Using Vector Base Amplitude Panning. *Journ. Audio Eng. Soc.* 45, 6 (1997), 456–466.
- Daniel Rudrich. 2018. Introducing the IEM Plug-in Suite. ‘klingt gut’ Intern. Symposium on Sound, Hamburg.
- Alia Sheikh, Andy Brown, Zillah Watson, and Michael Evans. 2016. Directing attention in 360-degree video. IBC 2016 Conference. <https://doi.org/10.1049/ibc.2016.0029>
- Meike Uhrig. 2015. *Darstellung, Rezeption und Wirkung von Emotionen im Film. Eine interdisziplinäre Studie*. Springer.
- Johannes Voit. 2014. *Klingende Raumkunst. Imaginäre, reale und virtuelle Räumlichkeit in der Neuen Musik nach 1950*. Tectum.
- Christiane Voss. 2009. Fiktionale Immersion. In *Es ist, als ob. Fiktionalität in Philosophie, Film- und Medienwissenschaft*, Gertrud Koch and Christiane Voss (Eds.). Wilhelm Fink, 127–138.
- Ka Yu, Matthew Brownell, Joslyn Schoemer, Daniel Neafus, Thomas Lucas, and Zachary Zager. 2007. Live Action Film Footage for an Astronomy Fulldome Show. *Planetarian* 36 (01 2007).
- Franz Zotter and Matthias Frank. 2012. All-Round Ambisonic Panning and Decoding. *Journal of the Audio Engineering Society* 60, 10 (2012), 807–820.







DOI: [10.22032/dbt.39936](https://doi.org/10.22032/dbt.39936)



URN: [urn:nbn:de:gbv:ilm1-2019200492](https://nbn-resolving.org/urn:nbn:de:gbv:ilm1-2019200492)