Master's Thesis

Perceptually Motivated Ambient Scene Recording and Parametric Embedding

Elisabeth Frauscher, BSc
Matr.Nr.: 01213371

Supervisor: DI Dr.rer.nat. Franz Zotter

Assessor: O.Univ.Prof. Mag.art. DI Dr.techn. Robert Höldrich

University of Music and Performing Arts Graz and Graz University of Technology

Master's Degree Programme: Electrical Engineering and Audio Engineering

Graz, April 21, 2020

Statutory declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Abstract

Various auditory experiments rely on faithful generation and reproduction of virtual audio scenes that yield a high reference to real acoustic environments. Most current approaches are based on simulated room impulse responses; however, highly complex and generic compositions are still challenging. The goal of the proposed master thesis is to develop a hybrid spatial recording technique that allows recording real 3-dimensional acoustic scenes and parametrized embedding of arbitrary source signals, whereas the latter is based on estimated sound field parameters derived from the recorded ambient sound scene. The perceptual quality of different spatial recording paradigms and source embedding methods will be evaluated by means of listening experiments. More specifically, dummy head recordings of chosen acoustic scenes will be compared to binaural renderings under consideration of loudspeaker-based reproduction methods. By placing loudspeakers in the recorded scene, reference embeddings will be used further on to validate the plausibility of the proposed approach.

Kurzfassung

Verschiedenste auf akustischen Reizen basierende Experimente erfordern eine originalgetreue Generierung und Reproduktion von virtuellen Audio-Szenen. Typischerweise werden dafür simulierte Raumimpulsantworten als Ansatz gewählt - komplexere generische Szenen stellen jedoch noch immer Schwierigkeiten dar. Das Ziel dieser Masterarbeit ist es eine hybride Aufnahmetechnik zu entwickeln, die ein realitätsnahes Aufnehmen 3-dimensionaler Szenen und das parametrische Einbetten verschiedenster Audiosignale in die jeweilige akustische Umgebung ermöglicht, wobei für Zweiteres aus den zugehörigen Aufnahmen Schallfeldparameter geschätzt werden. In Hörversuchen wird die perzeptive Qualität der verschiedenen räumlichen Aufnahme-Paradigmen und der Einbettung der Quellsignale evaluiert. Insbesondere werden Kunstkopf Aufnahmen der gewählten akustischen Szenen mit binauralen Renderings, unter Beachtung einer Lautsprecherbasierenden Wiedergabemethode, verglichen. Um die Plausibilität der vorgeschlagenen Konzepte zur Einbettung zu validieren, werden Lautsprecher in der aufgenommenen Umgebung als Referenz verwendet.

Contents

Introduction

To assess what people hear in everyday life, whether actively or subconsciously, and what they can still perceive if they have a hearing impairment, it is essential to carry out hearing tests. This requires specially built acoustic laboratories - like the one *sonible*, a hardware and software audio company from Graz, developed for a Canadian university. This project subsequently spans the framework of this master thesis in cooperation with *sonible*.

In order to meet the requirements of the *Brain and Mind Institute*, University of Western Ontario (UWO), a detailed and authentic reproduction of various everyday-life acoustic scenes is to be reproduced in the laboratory. Ideally, the reproduced scenes should be as generic as possible so that they can be flexibly adapted to the respective research topic under investigation. This means, for example, that it should be possible to embed sources adapted to the acoustic characteristics of the recording location, e.g. to integrate a speech recording automatically and plausibly into a recording of an office scene. The methods in the master thesis should be developed with a view to making it possible for research institutions to prepare new material themselves.

To enable a realistic acoustic image reproduction in the case of the laboratory at the UWO, *sonible* built a 91-channel loudspeaker dome (see figure 1). This allows sources to be distributed via direct routing in all directions - but it is also capable of playing back audio files up to 8th-order Ambisonics [ZF19a], which attempts to recreate a physically accurate and stable 3D-image of the sound field. As an interface a GUI (see figure 2) was developed, which supplies control data via OSC. Currently it offers the possibility to create loopable scenes, which are composed of an ambient file and sound objects positioned at certain distances and directions.

The topics of scene reproduction and plausible embedding of sound sources are currently explored in numerous studies. Many in the field of augmented reality and virtual reality, for example in the gaming and film industry, but most of them actually in the field of hearing-aid research [GEH15].

The focus is hereby mainly on room auralization, as with the LoRA (Loudspeaker-based Room Auralization) [FB12], the SOFE (Simulated Open Field Environment) [BSH10],

Figure 1 – 91-channel loudspeaker array in an unechoic room at the *Brain and Mind Institute*, University of Western Ontario designed and built by *sonible*

and the Virtual Reality System Aachen [DSP10], which uses the RAVEN framework. In other approaches, such as the binaural RAZR room acoustic simulator [TWE14], the focus is on accurate rendering of inter-aural and room acoustic parameters instead of correct physical reproduction of the sound field. Other studies, such as the open source tool TASCAR (Toolbox for Acoustic Scene Creation and Rendering) [GGH19], aim to increase ecological validity of the created virtual acoustic environments (VAEs). In case of TASCAR particular emphasis is placed on interactive and realistic modelling of dynamic sound sources such as Doppler shift and comb filtering.

All approaches presented above are essentially based on spatial acoustic modelling, e.g. by using the elaborate ODEON [1] software or by more restrained methods as the image source method (ISM) together with knowledge of absorption coefficients of the simulated room.

Simon, by contrast, who conducts experiments for hearing impaired peolpe at the Laboratory for Experimental Audiology in Zurich, relies mainly on HOA recordings as VAEs. For example, he uses Eigenmike recordings as background scenes and creates speech stimuli by convolution with HOA room impulse responses (RIRs) [SKWD19] [ND$^+$19]. Similarly, coincident microphone recordings are used by Minaar in his studies on the use of VAEs for the development of hearing aids [PM$^+$13].

In a round robin test [BAA$^+$19] state of the art of room acoustic modelling software

1. ODEON Room Acoustics Software, https://odeon.dk/

Figure 2 – GUI to create scenes and control playback of the loudspeaker dome.

both in the physical and perceptual realms were evaluated. The main findings were that authenticity is not achieved and that the methods examined are very complex. However, it is also acknowledged that authenticity is a very strict criterion, denoting that the simulation sounds exactly like the real room. Additionally it was stated that the observed differences between simulation and reference are mainly due to tonal colouration and sound source positions.

At this point it should be mentioned that in this thesis authenticity cannot be used as an evaluation criterion due to the lack of a real reference. Instead, plausibility is preferred, which implies that only relevant properties are reproduced and *"thus shifts from 'copying an existing environment in all it's physical aspects' to 'a suitable reproduction of all required quality features for a given specific application'"*, according to [Pel01, p.4]. In this context, the question of how much effort is necessary so that people cannot distinguish acoustic virtuality from reality arises.

Further evaluations in previous studies were performed to compare VAEs generated by room acoustic models with microphone recordings. In contrast to [OB16], who concluded that the results are independent of the environment used, it was shown in [AMD19] that rooms are better reproduced by RIRs measured with microphone arrays than by room acoustic simulations.

Overall, however, very few VAEs are based on recordings, nor that the embedding is carried out on the basis of a recording - as it is done in this study. The study finds its position, so to speak, between the two common methods of creating VAEs, since

both recordings and room acoustic parameters (estimated via the recording itself) are employed.

In terms of recording techniques, various spaced microphone arrangements have emerged over the last decades - typically as the main microphone for surround recordings [Wil10]. What these have in common is that they are always based on the same stereophonic principles [LJM17]. Lee has recently begun to study arrays in more detail that are independent of orientation and therefore better suited for ambient 3D-recording [SM09] [Lee19] [RL17].
Meanwhile, there are a lot of first-order Ambisonic microphones (occasionally including directional sharpening processing), as well as some Ambisonics microphones capable of recording in higher order [ZF19a] [Bat16] [SMB06].

The first chapter of my master thesis examines with which of those (3D-)recording techniques one comes closest to reality when listening to the recordings on headphones or in a 360° loudspeaker dome. Suitable microphone arrays will be pre-selected and a pilot evaluation test will be conducted. Reasonable acoustic scenes covering various places of everyday life are recorded to fill the ambient sound library for the psychoacoustic listening tests. This will require an understanding of what distinguishes a realistic image.

In the second chapter the chosen and implemented parameter estimation method is described. It is supposed to enable gathering of information about the present acoustics at an arbitrary recording location with a manageable effort.

By using those results as input data, the implementation of source embedding in a corresponding acoustic scene is presented. It will focus in particular on naturalness of the embedding and the reproduction of a plausible room impression regarding early and diffuse reverberation. It is aimed to answer the question of the extent to which the embedding has to go into detail in order to sound plausible.

Both the relevant microphone array candidates as well as the result of the implementation of parameter estimation and embedding will be evaluated in listening tests in chapter 4. A simulated office scene serves as a basis. For meaningful results it is necessary to define plausibility and find meaningful parameters for the evaluation of atmospheric scenes and embeddings.

Finally the results of this master thesis are presented and discussed in the conclusion.

Chapter 1

Plausible scene recording

Before starting to design a recording setup several key aspects in view of the target application need to be defined. First of all, recording ambient everyday life scenes requires a quite versatile microphone array that is able to cope in- and outdoor locations, various room sizes and hence different source distances, albeit mostly in the far field. Consequently, the array would ideally be robust toward weather conditions and other unforseeable incidents especially concerning people and should be unobtrusive not to modify the behavior within the recorded environment.

Generally, ambient sound broadly denotes rather distant environmental background sounds generated by traffic and machines, people, weather, animals and nature, present at the recorded location. A plausible perception also calls to capture the scene from the same perspective a human would normally hear, which corresponds to a center spot at a height of 1.65 m in Europe.

Regarding the 3D dome-like playback setup, a 360-degree sound field including height channels should be captured. To enable spatial processing of the scene the recording should be rotatable and therefore must not focus on a primarily frontal perspective, where the specific depth and direction mapping properties of spaced, coincident or Ambisonic main arrays need to be considered. Thus techniques appear favorable that produce a continuous phantom imaging, at least in the horizontal plane.

Also the anechoic room acoustic of the acoustic laboratory and the size of the sweet spot should be taken into account, especially for designing the evaluation test.

The above mentioned criteria disqualify some of common recording techniques. The following section candidates are being discussed together with their design tools.

1.1 Microphone arrays

The predominant pre-requisite the candidate microphone arrays must fulfill is to accomplish spatial auditory images of high plausibility [Lin15]. The requirements within this thesis also imply that they must be able to capture sufficient information to generate a valuable input for the automated parametric source embedding, whereas microphone arrays with a low signal-to-noise ratio (SNR) and high directional resolution offer an advantage. Ideally, the array would be portable, easy to set up and affordable. A relatively small size of the array adds another advantage for field recordings in narrow rooms as well as public spaces to prevent unnecessary conspicuousness - the less recording gear attracts people the more an unaffected scene will be recorded.

One can distinguish in the first place between two types of microphone arrays: coincident and spaced arrays. The former is ideally just determined by interchannel level difference (ICLD) and the latter both by interchannel time difference (ICTD) and ICLD.

Advantages and disadvantages of those categories are summed up in table 1.1. Through the higher channel separation and decorrelation of the microphone signals, spaced microphone arrays tend to achieve a better listener envelopment. Contrary, coincident arrays yield a good localization quality. The sound quality tends to be higher for spaced arrays, because the microphone selection is larger and space is not limited.

	Envelopment	Localization	Sound quality	Size, portability
Spaced arrays	+	-	(+)	-
Coincident arrays	-	+	(-)	+

Table 1.1 – Properties of spaced and coincident microphone arrays.

Until now microphone arrays were mostly designed for indoor frontal music performances with back and height channels capturing early and late reverberation. Comparative tests were additionally carried out with mainly music sources [LJ19].

In the following section possible microphone array candidates, grouped into the above mentioned two categories, will be considered more closely.

1.1.1 Spaced microphone arrays

Adjustable parameters when designing a spaced multichannel microphone array are number, type and directional characteristic of the microphone as well as spacing and angle between the cartridges, whereas none of those can be chosen independently. It was found in literature that most research so far was done on a correct image source localisation in the horizontal plane.

Image source localization is generally based on stereophonic localisation curves derived by an ICTD/ICLD trade-off relationship. Localisation curves desribe where a sound object is perceived in relation to where the target is positioned [SM09].

ICTD and ICLD are closely related to interaural level difference (ILD) and interaural time difference (ITD), which enable the localisation of sound sources for human hearing. The former by the level difference through the acoustic shadow of the head, which results in -20 to +20 dB (JND is 1 dB) and the latter by the difference of arriving time of a sound, which varies from -0.7 to +0.7 ms. Phase differences for low frequencies and the envelope differences for high frequencies determine the perception of ITD.

To shift the phantom image in a stereophonic setup a trade-off function for the ITD and ILD has to be defined. The most popular technique to shift the phantom image is amplitude panning by using the tangent law formulated by Clark and Dutton. It is a theoretical model to match the ITD of a phantom source with a real target source position and depends on the gains g_1 and g_2 of two loudspeakers located at $\pm\alpha$ with the source positioned at φ

$$\frac{\tan\varphi}{\tan\alpha} = \frac{g_1 - g_2}{g_1 + g_2}. \tag{1.1}$$

However, Pulkki and Karjalainen claimed that localisation judgement based on the tangent law does not work in the frequency region where human hearing is most sensitive, due to the fact that localisation relies upon ITD below 700 Hz but upon ILD above 2kHz. Their findings were confirmed by listening experiments conducted by Lee [LJM17].

Wittek and Theile [WT02] developed localisation curves based on the Williams curves, which were determined originally upon listening test data and polynomial interpolation. Williams moreover neglected any dependency of the SRA on loudspeaker base angle and source-array distance. In the localisation curves by Wittek and Theile a linear (13%/0,1ms and 7,5%/dB) trade-off within the 75% image shift region, and a logarithmic function for non-linear trade-off is assumed. In cooperation with Schoeps they designed a microphone array design tool based on their work called 'Image Assistant' [1] (see figure 1.1).

1. Image Assistant Schoeps, http://www.ima.schoeps.de/

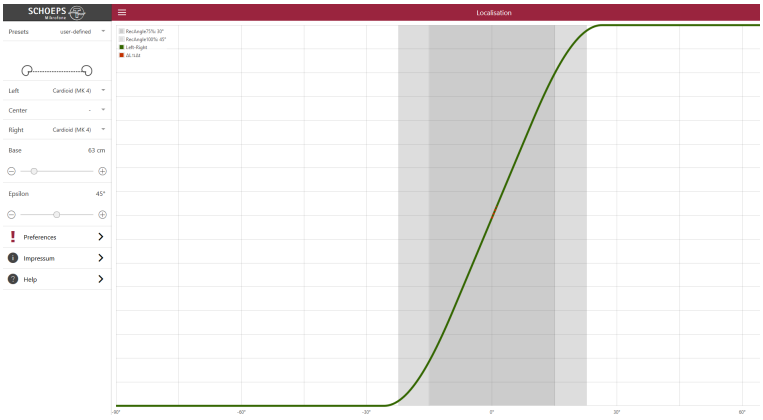


Figure 1.1 – Image Assistant example for a cardioid microphone pair with 45° angle. The recording angle fits the microphone angle at a base distance of 63 cm. Changing the source distance varies the recording angle from 40-46°.

Contrary to previous assumptions, Simon found that localisation curves vary depending on the position of loudspeaker segments, so that localisation curves derived for a frontal stereo loudspeaker pair are not applicable to loudspeaker pairs positioned to the side [SM09].

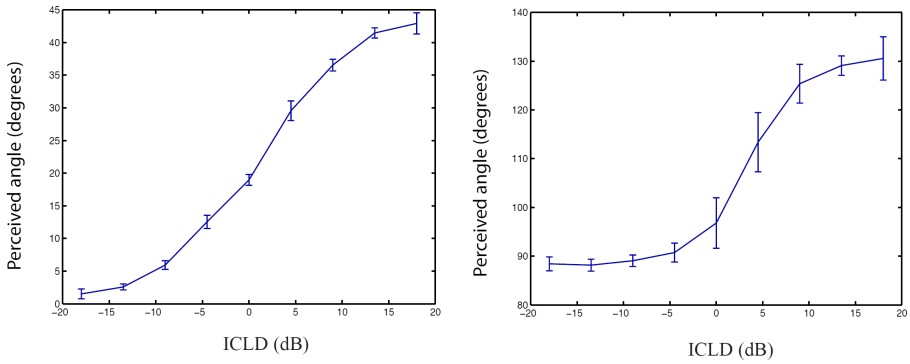


Figure 1.2 – Localisation curves for a front segment loudspeaker segment (left) and a segment located on the side (right) in dependence of the ICLD [SM09].

However, outcomes of listening experiments regarding localisation curves have been observed to diverge. A lot of those results could not be verified in comparison attempts at the IEM. Especially when participants were asked to move their head if the source position is ambiguous, answers were inconsistent. Zotter and Frank could confirm that localisation curves are more imprecise at lateral loudspeaker pairs, but do not appear to be steeper or even shifted to front [ZF19b].

In 2019 Lee conducted listening experiments to get further insight in stereophonic localisation curves on the side. Subsequently he developed a region adaptive ICTD/ICLD trade-off method together with a perceptually motivated amplitude panning (PMAP) law [LJM17]. Based on level differences instead of time differences it should produce a more stable and accurate panning function than the tangent law. The PMAP law also includes a perceptual scaling for an arbitrary loudspeaker base angle, which was computed based on a binaural ERB (equivalent rectangular bandwidth) model [LJM17].

Results of his listening tests showed that by using the PMAP law, the image source position agrees well on the target source position concerning ILD, but with slight ITD errors. The evidence was observed to be stronger at a source panning of 10-15°, which is explained by a higher localisation dependence on frequencies above 1 kHz, where ILD plays a more important role. Additionally, the scale factor for different loudspeaker base angles performed well. These findings were brought together by Lee in a sophisticated microphone array design tool called MARRS [2] (figure 1.3).

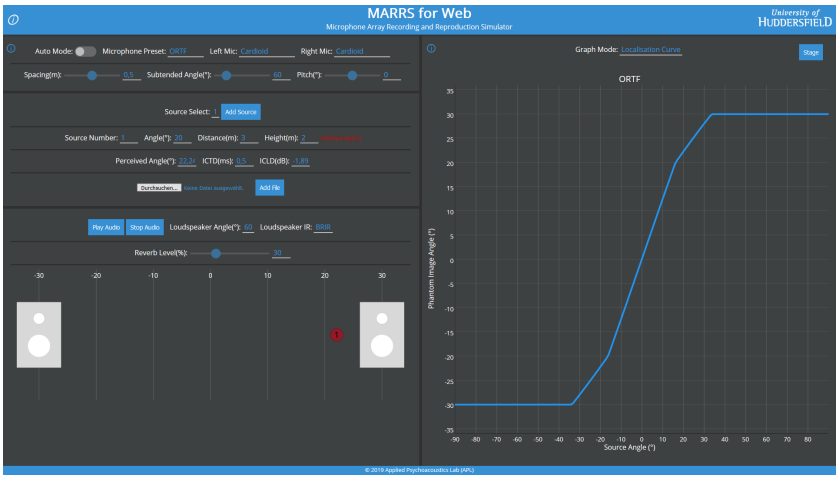


Figure 1.3 – Screenshot of the MARRS design tool [Lee19].

In contrast to the above discussed recording techniques the reproduction setup for playback is basically the same for all spaced microphone arrays, as all channels are routed (or panned, if vector-base amplitude panning (VBAP) is used for playback [ZF19a]) discretely to the equivalent loudspeakers without matrixing.

Generally, regarding previous studies, we expect spaced arrays to provide a great balance between localizability and spaciousness, due to the fact that they imply both types of aural cues, ICTD and ICLD, for phantom imaging [Lee19].

In order to derive possible spaced microphone array designs from this theory, a cross-section of models and approaches found in the literature is presented below.

2. MARRS design tool, http://marrsweb.hud.ac.uk

ORTF-3D Surround

A typical microphone array for ambient recording is the ORTF-3D Surround Microphone by Schoeps. It basically consists of two layers of a four-channel ORTF-Surround system including eight supercardioids with a 10 or 20 cm spacing - in other words four vertical XY pairs.

Figure 1.4 – Exemplary mounting of an ORTF-3D microphone array. [3]

Since we are used to listening to stereophonic reproduction setups, the ORTF-3D array is perceived to sound very naturally although it does not attempt recreating the original sound field. Due to the elliptical shape of the array the surround recording is not perfectly rotatable without losing locatability of sound sources.

Advantages on the other side are potential high quality of the individual microphones and the compactness of the array, which fits easily in a weather-resistant windshield. The channel crosstalk is hereby minimized by the use of supercardioids.

IRT Cross (2D) and Hamasaki Cube (3D)

In contrast, the IRT ('Institut für Rundfunktechnik') Cross provides four equally distributed cardioid microphones positioned in a square with 25 cm spacing between each microphone. It was originally designed for the purpose of recording diffuse sound in setups with a main stereophonic array.

3. ORTF 3D, https://schoeps.de/en/products/surround-3d/ortf-3d.html

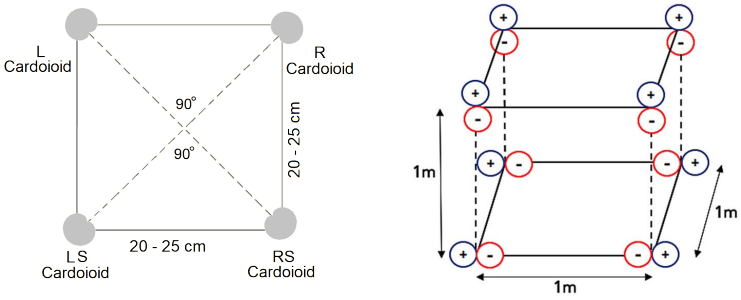


Figure 1.5 – Microphone array schemes of the IRT Cross [6](left) and the Hamasaki Cube [7](right).

Another convenient equally spaced design, which would additionally compensate the missing height layer, is the Hamasaki Cube. The array uses four figure-of-eight microphones in a distance of 1 m with opposite orientation on the bottom layer. Four further figure-of-eight microphones with the null pointing towards the sides are placed on the top layer, in order to capture diffuse sound from above and below.

In the following, an array that attempts to combine advantages of the introduced convenient spaced microphone arrays is presented.

ESMA (Equally Segmented Microphone Array) and ESMA-3D

The concept of critical linking, which was already proposed in 1999 by Williams and Dû [WD99], defines the setup of an ESMA. If the stereophonic recording angle (SRA) of each circular arranged microphone pair links seamlessly to 360° without overlap, or in other words if the microphone and recording angles are equal, a correct image shift can be achieved.

The main criterion here is the choice of an accurate microphone distance in conjunction with the segment angle of the microphone pairs. Following the discussion in the introduction this has to be taken into account as different base angles produce different localisation curves.

Williams originally introduced a 5-channel horizontal ESMA and called an extension up to 6-8 channels due to upcoming 7.1 reproduction setups [Wil10]. Based on his developed localisation curves for several direction patterns he proposed a spacing of 39 cm for his 5-channel design. Simon [SM09] conducted experiments on localisation curves for an octagonal microphone and loudspeaker setup for the principal purpose of

7. IRT Cross, https://schoeps.de/en/products/surround-3d/ortf-surround-irt-cross/irt-cross-set.html
8. Hamasaki Cube, https://www.abbeyroad.com/

designing 360° microphone arrays.

Lee conversely did further research on a quadraphonic array with a SRA of 90-degree [LJM17] and conducted listening experiments in order to compare various microphone spacings.

Lee hereby resumed on phantom image localization accuracies that the 50 cm microphone spacing produced the best phantom imaging and that the closer a sound source is positioned on-axis the more stable is the localization. Furthermore he showed that the PMAP law and the proposed scaling method, integrated in the MARRS design tool, produced more accurately panned images than the tangent law. [Lee19]

In line with Williams, Sena and Simon, also Lee concluded that finer ESMA resolutions should improve the unstable phantom centre image localization in sound field rotations. For this reason my subsequent study involves a 6 and 8-channel ESMA (due to symmetry).

For the two chosen ESMA configurations a comparison of suggested microphone spacings d_{mic} and resulting ring radii r by the MARRS App and the Image Assistant is given in table 1.2.

	Base angle	MARRS App d_{mic}	MARRS App r	Image Assistant d_{mic}	Image Assistant r
ESMA6	60°	0.5 m	0.5 m	0.39 m	0.39 m
ESMA8	45°	0.55 m	0.72 m	0.63 m	0.82 m

Table 1.2 – Microphone spacings and ring diameters for three selected ESMA configurations as suggested by the MARRS App and the Image Assistant.

An examination on the recommended spacings can be done by using the $\boldsymbol{r}_E$-vector model. Commonly the model is applied for predicting perceived direction in loudspeaker setups as investigated in [Fra13]. In this thesis it will be used to specify how uniformly directions are mapped by various ESMA configurations and how wide the direction mapping is.

The general $\boldsymbol{r}_E$-vector model is denoted as

$$\boldsymbol{r}_E = \frac{\sum_{m=1}^{M} g_m^2 \boldsymbol{\theta}_m}{\sum_{m=1}^{M} g_m^2}, \tag{1.2}$$

whereby g_m is the amplitude gain of the m-th microphone and $\boldsymbol{\theta}_m$ is defined as the unit vector of the m-th microphone orientation. Due to the normalization the magnitude of $||\boldsymbol{r}_E||$ is set to the value range between 0 (no direction) and 1 (only one distinct direction).

The model can be extended for source positions at different angles and distances [ZF19a] [WF18]. For this purpose, an additional distance-dependent sound absorption weight modelling the -6dB damping per distance doubling is introduced

$$w_{d,m} = \frac{1}{||\boldsymbol{r}_{\theta_{1..m,\boldsymbol{v}}}||}, \tag{1.3}$$

$||\boldsymbol{r}_{\theta_{1..m,\boldsymbol{v}}}||$ is hereby the length of the sound path to each microphone m.

Moreover the law of the first wavefront is considered by a time-dependent weight, which is applying less weight to microphones regarding the arriving time delay τ_m in s

$$w_{\tau,m} = 10^{\frac{1000 \cdot w_\tau \cdot \tau_m}{20}}. \tag{1.4}$$

As investigated in [Kur18] and [Weg20], attenuations derived from the echo threshold w_τ = -0.1 $\frac{\text{dB}}{\text{ms}}$ for static sounds, w_τ = -0.5 $\frac{\text{dB}}{\text{ms}}$ for transients and w_τ = -0.25 $\frac{\text{dB}}{\text{ms}}$ for a generally valid compromise have been successfully applied for localization prediction.

The extended $\boldsymbol{r}_E$-vector model is thus denoted for broadband sound sources as

$$\boldsymbol{r}_E = \frac{\sum_{m=1}^{M}(w_{d,m} w_{\tau,m} g_m)^2 \boldsymbol{\theta}_m}{\sum_{m=1}^{M}(w_{d,m} w_{\tau,m} g_m)^2}. \tag{1.5}$$

Following Frank's experiments [Fra13] listeners tend to hear a source width of $\frac{5}{8}$ of the aperture angle spanned between loudspeaker pairs, so that the source width can be expressed by

$$W = \frac{5}{8} \cdot \frac{180°}{\pi} \cdot 2 \cdot \arccos ||\boldsymbol{r}_E||. \tag{1.6}$$

The results obtained from the above presented model are now compared regarding the $\boldsymbol{r}_E$-vector angle error and the source width W. The ring radii determined from the MARRS App, the Image Assistant and additionally a resulting mean value of 1.24 m are included in the comparison (table 1.2). Furthermore, the results for ESMA6 and ESMA8 are analysed at different source distances (2m, 3m, 4.5m) and for the above presented echo thresholds w_τ.

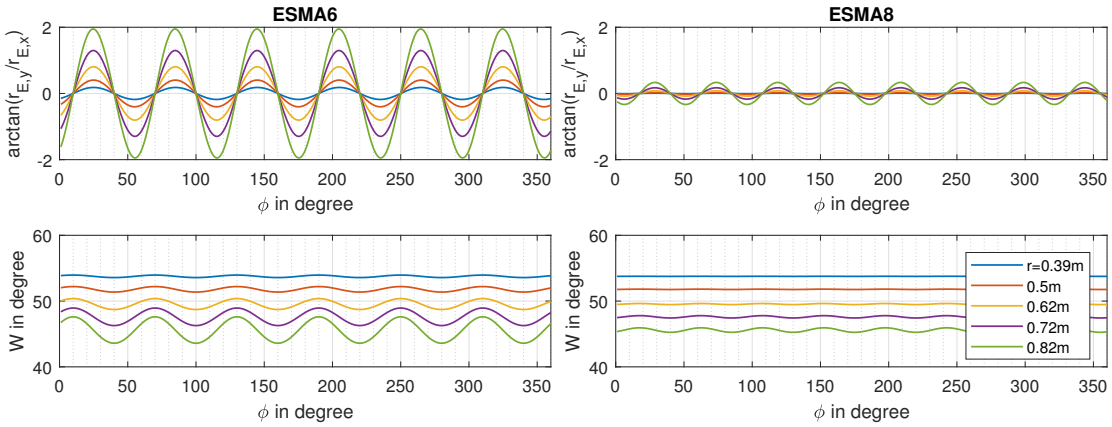


Figure 1.6 – $\boldsymbol{r}_E$-vector angle error and source width W shown for several radius configurations of both ESMA6 and ESMA8. $w_\tau = -0.25 \quad \frac{dB}{ms}$ and source distance d = 2 m.

It is inherent that the source width increases and the angular amplitude decreases with distance (figure 1.6-1.8). The closer a source is to the array the more it will move from one loudspeaker to the other during playback. Contrary, the more microphones are used (see comparisons between ESMA6 and ESMA8), the smaller the angular amplitude difference becomes. For W, on the other hand, only the ripple decreases as the number of channels increases.

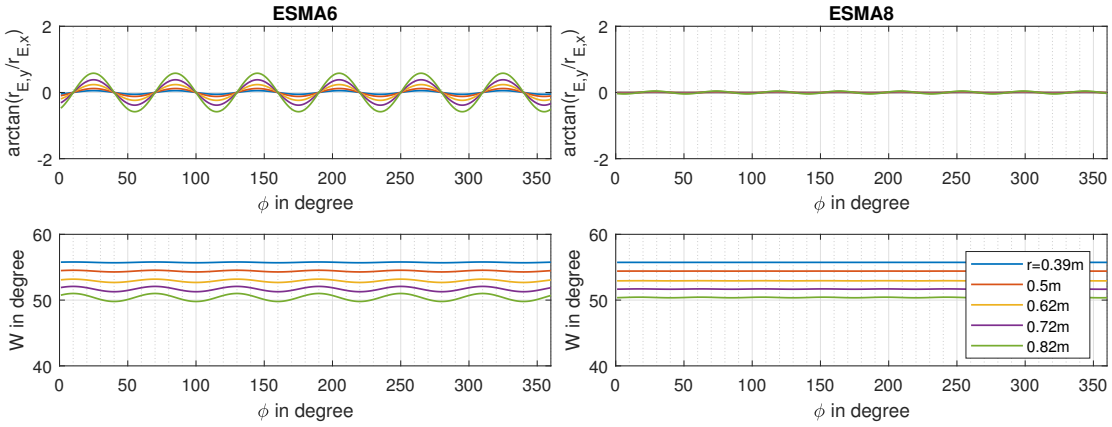


Figure 1.7 – $\boldsymbol{r}_E$-vector angle error and source width W shown for several radius configurations of both ESMA6 and ESMA8. $w_\tau = -0.25 \quad \frac{dB}{ms}$ and source distance d_5 = 3 m.

Moreover, as confirmed in figure 1.6 and figure 1.7, a smaller ESMA ring radius leads to a lower ripple of the $\boldsymbol{r}_E$-vector angle error, but simultaneously to a higher source width and vice versa.

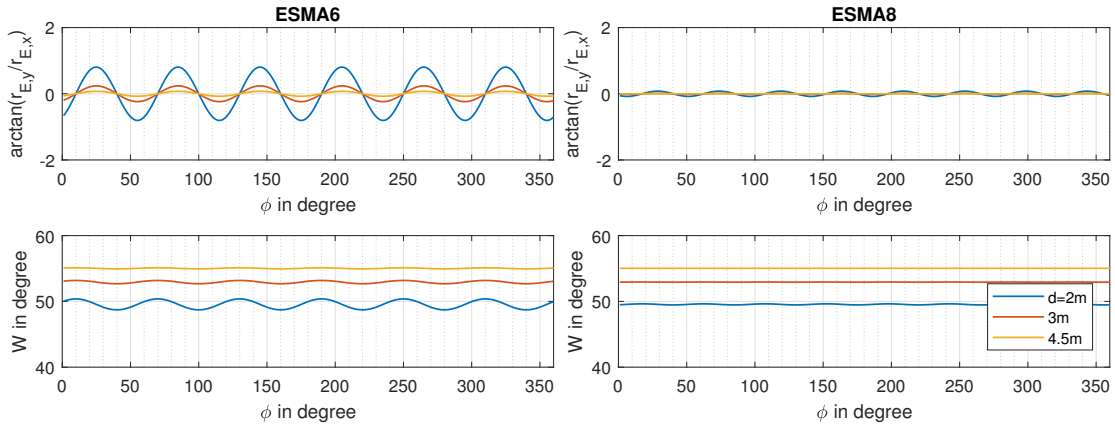


Figure 1.8 – $\boldsymbol{r}_E$-vector angle error and source width W shown at various source distances d for both ESMA6 and ESMA8. r = 0.62 m and $w_\tau = -0.25\frac{dB}{ms}$.

Inspection of the comparison in figure 1.9 confirms that transient sounds ($w_\tau = -0.5\frac{dB}{ms}$) lead to a lower perceived source width.

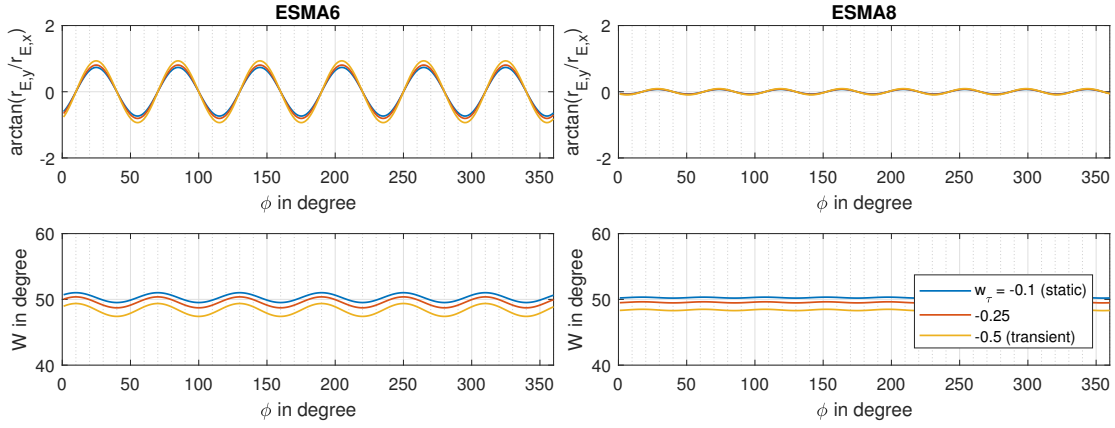


Figure 1.9 – $\boldsymbol{r}_E$-vector angle error and source width W shown for various values of w_τ for both ESMA6 and ESMA8. r = 0.62 m and ESMA to source distance d = 2 m.

For the investigated aspects - angle error and source width - better results are achieved for the ESMA8. However, apart from the higher ripple of W for the ESMA6, both configurations show the same weak source width image. Although the value range of

the $\boldsymbol{r}_E$-vector angle error seems to be negligible in comparison to that, the JND can cause a perceptual difference. For frontal directional resolution of hearing, a JND of around 1° is reported in literature [Bla99]. In contrast to the ESMA8, the r_E-vector angle error for the ESMA6 would thus be above the JND from a ring radius of 0.72m and up to a distance of 2m. At least for directional localisation, the higher channel number of an ESMA8 would therefore only provide a relevant advantage for recordings of sound sources closer than 2m, according to the model. From a distance of 4.5m the number of channels does not seem to have an impact anymore.

Since the ripple amount of the angle error is inverse to W, the optimal ring radius could be defined where the ripple is still within the JND and W is minimal at the same time. For the ESMA6 this would lead to a radius of 0.62m. In case of the ESMA8, the source width could be reduced if a larger radius is selected. However, for direct comparability of the recordings and better transportability, which would suffer from a larger diameter, the same radius was employed for the ESMA8.

To capture height and room information upwards facing microphones complete the ESMA-3D. While Williams proposed to use cardioid microphones [Wil13], Wittek and Theile suggested figure-of-eigth capsules with the null pointing towards the main sound source. However, ground reflections shouldn't be represented that prominently in the recording. Lee consequently recommended the use of supercardioids, especially for the case when sound events occur at the same height level of the array, as they provide a level attenuation of -10dB for sound arriving from 90°.

Generally, it seems that for vertical panning spectral cues overrule inter-aural cues [ZF19a]. In addition, time differences and decorrelation are not as effective as on the horizontal plane - moreover ICTD is rather unstable in the vertical layer. To examine the effect of microphone layer spacing in case of the ESMA, Lee conducted listening experiments with regard to vertical localization, spatial impression and preference [LG14]. Considering 0m, 0.5m, 1m and 1.5m distances, he concluded that a vertical distance has no significant influence on the perceived spatial impression. The coincident layer was rated with a higher preference probably due to less comb filter effect at the listener position. This advantageous result implies a more subtle and faster mounting. Regarding the level relation Lee suggested an ICLD of 7-9 dB, synonymous with mixing the height channel by this factor lower than the channels on the horizontal ring. This ICLD should avoid an upward shift of sound sources on the horizon.

Based on own, informal listening experiments and the recommendation of ITU-R for 'Advanced sound system for programme production' [Int18, p.13], a 30° elevation angle for the height channels is chosen concerning the reproduction of an ESMA-3D recording. What is more, since the radius of a sphere at this height is smaller it can be argued that fewer microphone channels for the same angular resolution can be used than on the horizontal plane. For this reason only 4 height-layer channels with 0 cm vertical spacing

will be used. The ESMA arrays including the height layer will further on be denoted as ESMA6h4 and ESMA8h4.

Regarding mounting of the ESMA-3D, a wooden ring with a diameter of $1m$ and drilled holes at the correct positions for the microphones is used. Small stereo bars were used to fix the cardioid and supercardioid capsules at the same position (see Figure 1.10).

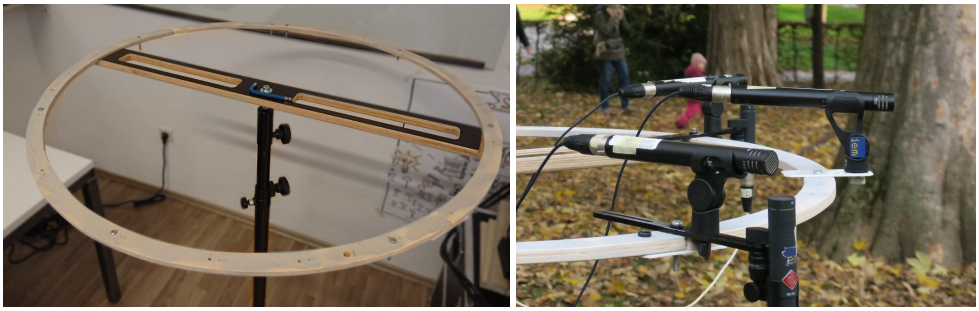


Figure 1.10 – ESMA ring construction (left) and microphone mounting (right).

In contrast to suggestions to place ESMA setups at a height of 2 m, as typically used for music recordings, the array should be placed at ear height in case of perceptionally plausible ambient sound recordings.

The benefits of this array according to previous studies are

— a balanced sound and low noise level due to the selected microphones,
— good externalisation qualities,
— a big sweet spot, which makes listening in the loudspeaker dome more comfortable because head movements won't change the sound, and
— plausible imaging of distance.

On the other hand, however, the amount of needed equipment is relatively high, transportation and the process of mounting are rather elaborate, and the contraption is rather large. Additionally, the total price of approximately 14.500-17.400 € depending on the ESMA configuration (calculated based on Schoeps microphones, excluding wiring and mounting) is relatively high compared to the other microphone arrays.

Another disadvantage is the difficulty to manage independence of power supply due to the number of channels. Field recorders generally support a maximum of 8 channels, consequently two *Zoom F8* field recorders would be needed. By way of illustration, both need to be filled with 8 AA batteries for approximately 30 minutes of field recording time.

A comparison of costs of all above introduced spaced microphone arrays shall be given in table 1.3. To maintain qualitative comparability the calculation was performed by assuming Schoeps microphones for each array.

	Count	Item	Estimated price
ORTF-3D	8x	CCM 41	13.520 €
IRT-Cross	4x	CCM 41	6.760 €
Hamasaki Cube	8x	MK 8G	13.528 €
ESMA6h4	6x	MK 4G Cardioid	8.670 €
	4x	MK 41 G Super Cardioid	5.920 €
			14.590 €
ESMA8h4	8x	MK 4G Cardioid	11.560 €
	4x	MK 41 G Super Cardioid	5.920 €
			17.480 €

Table 1.3 – Comparison of total microphone costs for the above introduced spaced arrays.

1.1.2 Coincident arrays

Coincident arrays got popular among sound recordists with the rise of VR content production and 360° video. The Ambisonics 3D-audio format enables flexible spatial processing as well as rendering to various reproduction setups. The typical compact microphone is easy to set up, fits also in small venues and can even be mounted on 360° cameras.

The mapping properties of the coincident arrays can be analysed analogous to the ESMAs by using the $\boldsymbol{r}_E$-vector model, which was introduced in the previous section. With respect to a recorded source direction $\boldsymbol{\theta}_s$ and the order N, Ambisonics microphones deliver an ideal image $\boldsymbol{r}_E = ||\boldsymbol{r}_E||\boldsymbol{\theta}_s$ with direction-independent width. The smallest width is achieved with the max-$\boldsymbol{r}_E$ image, which corresponds to $max||\boldsymbol{r}_{E(2D)}|| = \cos(\frac{90°}{N+1})$ for each pure horizontal (2D) reproduction and $max||\boldsymbol{r}_{E(3D)}|| = \cos(\frac{137.9°}{N+1.51})$ with a vertical plane (3D). This results in a source width of W =115.2 ° (eq. 1.1.1) for an order N=1 and for higher orders, for example N=3, W= 60.12° (eq. 1.1.1). Furthermore, in retrospect, this can also explain the upper limit of the source width W in figure 1.7 and 1.8 for the ESMAs. For a distant source and stationary signals (no precedence effect), this results in $W = \frac{5}{8} \cdot 2 \cdot \arccos(cos(\frac{90°}{2})) = 55°$.

Common coincident arrays, distinguished in first-order-Ambisonics (FOA), higher-order Ambisonics (HOA) and mixed-order microphones, will be presented below.

First-order Ambisonics (FOA) microphones

Since the first FOA microphone was developed by Gerzon in the 1970s, a number of further microphones differing in price, capsule quality and construction entered the market until today. Typically cardioid membranes are arranged tetrahedrally, resulting in a 4 channel B-Format and perfectly matching the channel amount of a convenient field recorder.

SoundField ST450 MK II

The SoundField ST450 microphone was pre-selected following a FOA microphone comparison conducted by Kurz et al. [EKF15]. The arguably most popular Ambisonics microphone on the market received the highest ratings especially regarding envelopment [9], but with the addition that it is also the most expensive one (see comparison in table 1.4).

Figure 1.11 – SoundField ST450 MKII microphone and control unit [12]and detail view of the 4 membranes inside the housing [13].

Enclosed are four half-inch cartridges positioned at a smaller microphone spacing compared to other FOA microphones.

It was noted in [EKF15] that the standard rotation of the microphone yields better sound localization results than in the 90-degree rotated mounting. A possible reason could be that the two forward looking capsules reproduce the most familiar stereo panorama.

9. The SoundField ST450 was compared to the Oktava MK4012 and the SoundField SPS200.

13. SoundField ST450 MKII, https://www.steller-online.de/Audiohardware/Mikrofone/Soundfield/Soundfield-ST450-MK2-Kit-1.html

14. https://www.teltec.de/out/media/ST450-MKII-manual.pdf

Sennheiser AMBEO VR

Based on the design of the SoundField MKV, a FOA microphone alternative was release by Sennheiser in 2016 (figure 1.12).

In a comparative study the AMBEO VR was found to perform *"[..] very well in terms of directionality, on par with the Eigenmike, and significantly better for elevation accuracy compared to the Soundfield MKV."* [EB+17, p.9]. The timbre on the other side was perceived to be brighter than ideal. An 'Ambisonics Correction Filter' incorporated by default within Sennheiser's A-to-B-format conversion plugin, seems to apply a significant boost above 10 kHz.

ZOOM H3-VR

The ZOOM H3-VR is a lowbudget FOA microphone that perfectly suits for field recordings, since it already incorporates a recording device and an USB power port. Further more it offers wireless controlling and can be mounted easily with a total size of 7,6 x 7,8 x 12,3 cm. The microphone was not detected as candidate at the beginning of this study, but it is assumed to produce similar qualitative results as the SoundField SPS200.

Figure 1.12 – Sennheiser AMBEO VR microphone [17](left). ZOOM H3-VR microphone and field recorder [18](right).

However, in practice the ideal spatial reconstruction assumptions cannot be met for FOA microphones. This is due to diffraction effects and spatial aliasing, which occurs at lower frequencies for larger capsule distances. FOA microphones generally lack in

18. AMBEO VR, https://www.soundandrecording.de/equipment/sennheiser-ambeo-vr-mic-3d-audio-mikrofon/

19. ZOOM H3-VR, https://www.zoom-na.com/products/field-video-recording/field-recording/zoom-h3-vr-handy-recorder

inter-channel signal separation and poor envelopment for diffuse sounds due to their lack of time and limited level differences.

Carrying out parametric audio processing on FOA recordings could be used to obtain a resolution enhancement, e.g. by DirAC [GS16], COMPASS [20] or HARPEX [21]. The latter is chosen to be part of further examination.

Berge's patented HARPEX method (high angular-resolution plane-wave expansion) is accomplished by rotationally aligning the tetrahedral first-order layout with sources detected in every subband [Ber19]. Within the plugin several Ambisonics input formats and output layouts can be chosen. The processing can be edited by controlling rotation, filter settings and the amount of envelopment (ENV = 0-2).

Figure 1.13 – Screenshot of an exemplaric HARPEX plugin setting.

Higher-order Ambisonics (HOA) microphones

HOA microphones overperform first-order Ambisonics recordings amongst others regarding localisability and locatedness, a larger sweetspot area and the sharpness of the reproduced image.

em32 Eigenmike

The spherical Eigenmike EM32 produced by MH Acoustics was the first commercially available HOA microphone generating a 4th order Ambisonics signal. 32 omnidirectional electret pressure capsules are positioned on a 8.4 cm sphere. The microphone

20. COMPASS, http://research.spa.aalto.fi/projects/compass_vsts/plugins.html
21. HARPEX plugin, https://harpex.net

signals are transmitted to the Eigenmike Microphone Interface Box (EMIB) and calibrated and converted to Ambisonics via an EigenStudio plugin.

For a detailed analysis of the microphone, reference is made to [ZF19a]. Measured directivity properties of the microphone are shown exemplary in figure 1.1.2.

An open-access database of ambient Eigenmike recordings was released by Green and Murphy [GM17], which is referenced in terms of tonal characteristics.

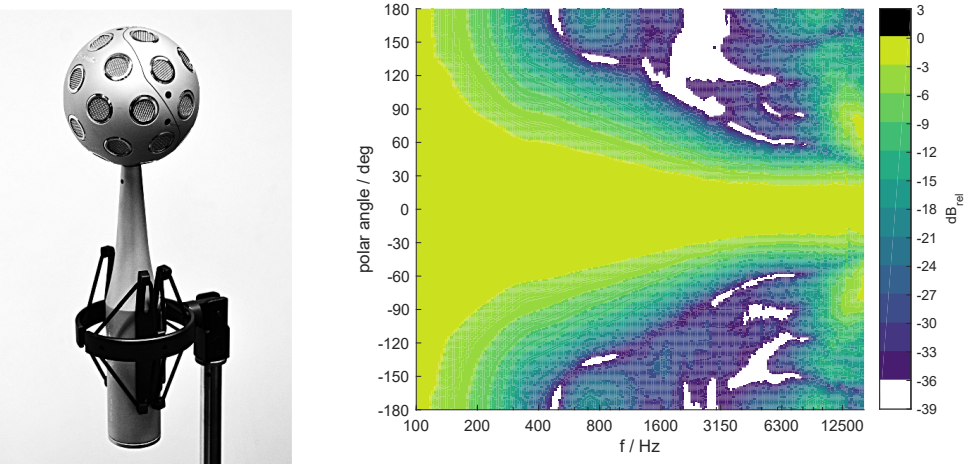


Figure 1.14 – Left: em32 Eigenmike 32-channel HOA microphone by MH Acoustics [24]. Right: Directivity surface plot for the em32 Eigenmike for $\chi = 90°$ and a beam pointing towards 0° elevation and azimuth (10dB noise boost). The plot was created by using the balon_holo visualization tool [25] implemented in MATLAB.

Zylia ZM-1

Recently a lower cost and more convenient to set up HOA microphone came on the market. The Zylia contains 19 MEMS capsules (69 dB SNR and 24-bit resolution) and measures 15,5 cm x 10,3 cm. Advantages of the utilized MEMS capsules (MicroElectro-Mechanical Systems) are the small size, which is about 10 times smaller than electret cartridges.

Regarding the number of channels, the recording can be encoded to a 3rd-order Ambisonics signal. Since the microphone output is USB, no interface or additional recording gear to a laptop is needed. To provide independence from a recording software running on a laptop, Zylia also released an USB field recorder.

25. em32 Eigenmike, https://link.springer.com/chapter/10.1007/978-3-030-17207-7_6
26. balon_holo visualisation tool, https://git.iem.at/p2774/balloon_holo

Figure 1.15 – Picture of Zylia ZM-1 microphone [29](left) and the Ambisonics converter plugin [30](right).

A comparison of the directivity plots of em32 Eigenmike and Zylia ZM-1 (figure 1.1.2 and 1.1.2) reveals only a few differences. In case of the Zylia ZM-1 more side lobes and spatial aliasing effects seem to arise at higher frequencies compared to the em32 Eigenmike.

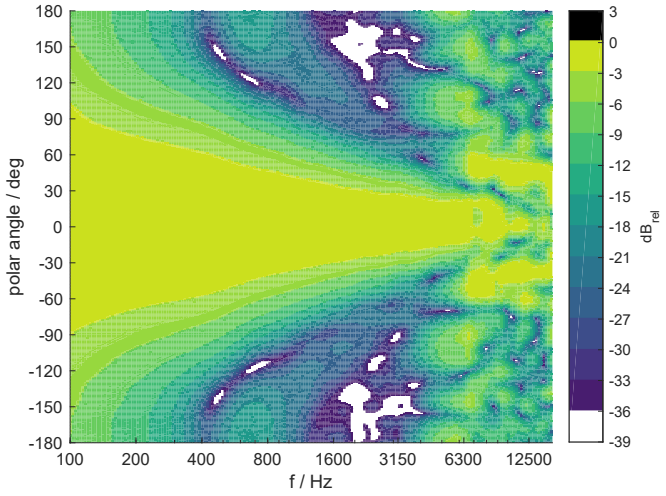


Figure 1.16 – Directivity surface plot for the Zylia ZM-1 for $\chi = 90°$ and a beam pointing towards 0° elevation and azimuth (10dB noise boost).

30. Zylia ZM-1, https://www.bonedo.de/artikel/einzelansicht/zylia-portable-recording-studio-test.html
30. Zylia Plugin, https://www.zylia.co/zylia-ambisonics-converter-plugin-download.html

Generally, HOA microphones seem to suggest a trade-off between audio quality and directionality. In the Ambisonic microphone comparison test conducted by Bates [Bat16], which included the em32 Eigenmike among several FOA microphones, the hypothesis that a HOA microphone would produce the best results regarding directional accuracy could be confirmed. However, artefacts and noise together with a dull perception of timbre led to a low rating in comparison to the dummy head reference and the other microphones.

As supplementary information, a comparison in the form of a cost table is offered for the presented FOA and HOA microphones in table 1.4.

	Estimated price
SoundField ST450 MK II	6.750 €
Sennheiser AMBEO VR	1.524 €
ZOOM H3-VR	289 €
HARPEX plugin	589 €
em32 Eigenmike	30.000 €
Zylia ZM-1	1.222 €

Table 1.4 – Comparison of total microphone costs for the above introduced coincident arrays.

Mixed-order microphones

An alternative to compensate for the respective disadvantages could be mixed-order microphones, whereas none has captured the market until now. Reference is made here to the dissertation by Marshall, who examined the application of the mixed-order Ambisonics method for microphone arrays in terms of performance and sound-field reconstruction [Mar14].

MARVIN (Microphone Array for Realtime and Versatile Interpolation)

This boundary layer microphone [MWK12] consists of a total of eight channels - 6 on the horizontal plane and one on each pole as shown in figure 1.17. Thus, horizontal microphone signals can be encoded in higher-order, where the human directional resolution is higher. Moreover, the 20cm diameter of the microphone offers a good prerequisite for binaural decoding and playback due to its similarity to the head dimension.

The microphone is still in development but could offer an interesting possibility to combine all needs of an ambient recording microphone array, in between spaced and coincident arrays and with space for studio-quality microphones.

Figure 1.17 – MARVIN mixed-order microphone [MWK12].

1.2 Pilot test

The above presented microphone arrays will be evaluated in detail in two listening experiments in section 4.2. Before the main evaluation a pilot test was conducted, whose results will be analysed in the following. It is referred to section 4.2 regarding all specifications and test design decisions, since all listening tests are based on the same samples (subsection 4.2.1) and statistical methods (section 4.1).

In a listening test, coincident and spaced arrays as well as high quality and lower cost arrays are to be compared against each other. A further test condition is the comparison of FOA and HOA microphones. Including a dummy head reference, nine arrays or configurations were selected for the binaural pilot test:

— KU100
The binaural dummy head microphone was used as reference and ground truth of naturalness.

— ESMA and ESMA-3D
In the category of Spaced Arrays, the ESMA-3D is the one we have high expectations of because it combines studio quality and natural sounding cardioids with high spatial resolution. The listening test should show whether a difference between ESMA6h4 and ESMA8h4 is perceptible. The corresponding ESMA configurations without a height layer were included to investigate the effect of the vertical layer.

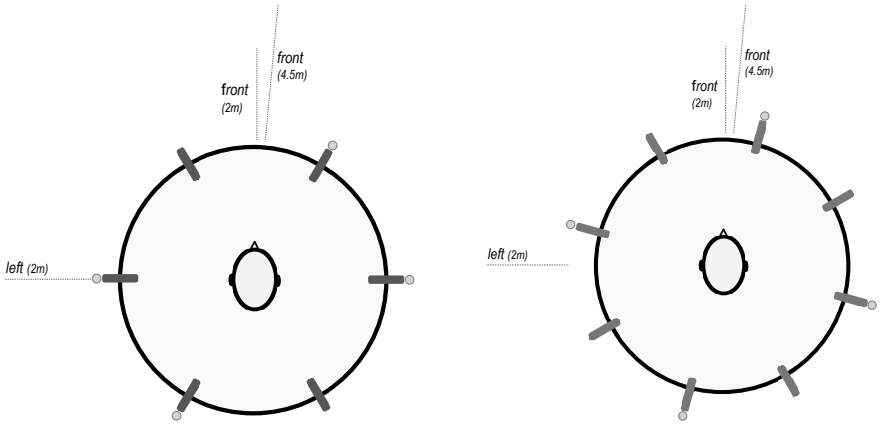


Figure 1.18 – Schemes of the ESMA6h4 (left) and ESMA8h4 (right) with depicted direction of sound sources used as stimuli in the listening experiment. Light grey circles depict upward facing supercardioids.

— SoundField ST450 MK II
For the coincident arrays, the SoundField ST450 was chosen as it stands for a wide range of different first-order Ambisonics microphones (Zoom H3-VR, Rode NTSF-1, AMBEO VR, SPS200, CoreSound Tetramic).

— ST450 HARPEX-upmix
The directionally sharpened version of the first-order SoundField ST450 microphone is included to explore the capacities of parametric upmixing algorithms. Two different plugin parameter settings regarding the amount of envelopment were investigated. ENV1 denotes a neutral envelope and ENV2 a maximum envelope portion.

— Zylia ZM-1
The simplicity and price of the Zylia microphone are compelling, so that it is included instead of the higher quality, but at the same time already more extensively analysed Eigenmike EM32.

The stimuli are referred to in the following as KU100, ESMA6, ESMA6h4, ESMA8, ESMA8h4, ST450, ST450 Harpex (ENV1), ST450 Harpex (ENV2) and Zylia.

In the pilot test 14 experienced listeners took part. The participants were between 22 and 58 years old, with an average of 29 years. Due to the online availability of the test (see description in the evaluation chapter in section 4.2.3) they could decide for themselves whether they would like to participate from home or not. The test was divided in two

parts, each containing the same three trials which were two times an office background scene including a speaker located in the front and one time an office ambience without additional speaker. In the first part the subjects were asked to rate the naturalness of the presented stimuli and in the second part they should rate the same stimuli in comparison to a dummy head reference. On average, the time required for the listening test was 50 minutes including individual short breaks. The tasks were originally given in German, but herein translated to English:

Naturalness, Part 1

> *Please rate the naturalness of the binaural snippets.*
> *How well do you get the feeling of 'beeing there'?*

The participants were supplied with a rough sketch of the room geometry and some sound source positions to get familiar with the office recording location. Additionally they got questions to frame a definition of naturalness:

> *How well can the directions be recognized and separated?*
> *Is it possible to estimate from the recording how the room is acoustically conditioned?*
> *Can differences in distance between the sources be perceived?*
> *How natural does the distribution in the spectrum sound?*
> *How good is the tonal quality of the recordings?*
> *How well does the wrapping work or does it feel being present in a 360° sound panorama?*

Similarity, Part 2

> *This time a reference representing the ground truth of naturalness will be available. Your task in this part: Rate the similarity between the reference and the other audio snippets.*
>
> *Please DO NOT include slight variances in the reproduced directions of 'Speaker 1' (in the front) in your evaluation.*

Regarding the results of the pilot test it should be noted, that two participants were excluded during the post-screening process due to inconsistency. Furthermore, the results for the Zylia microphone are out of competition because of a mistake in handling its signals. Since the Zylia converter plugin already provides SN3D normalised signals with side-lobes supressed, the wrong normalisation caused a playback of effective order that was roughly first. Therefore the Zylia results are labelled by 'corrupted'.

In the first part the dummy head recording was rated to be significantly the most natural one. It was confirmed to be reasonable to use the KU100 as reference for the second part. Generally the spread of given answers is rather large for the first part. It was observed that the missing reference led to a subjective interpretation of the scale. The presentation of a reference in the second part enabled a better comparative judgement and thus smaller variances.

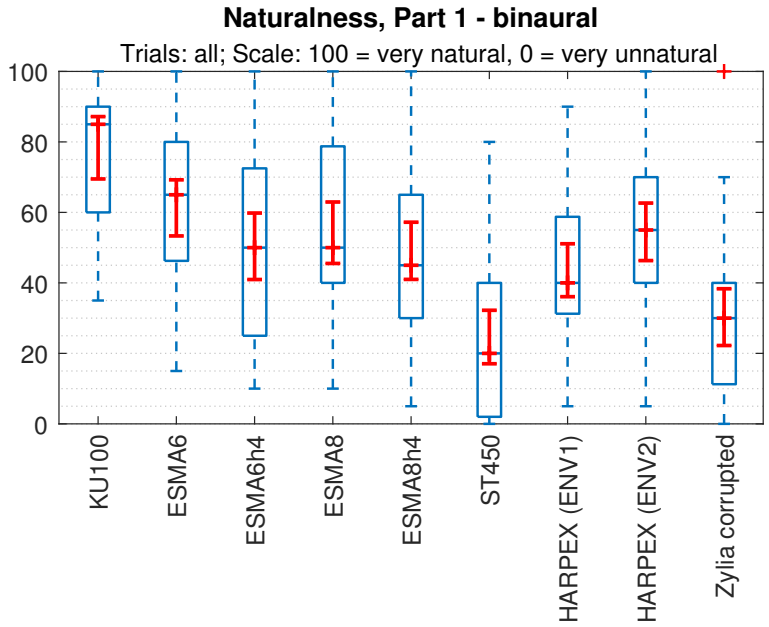


Figure 1.19 – Perceived naturalness in Part 1 of the binaural pilot test without a presented reference. Blue boxes show the inter-quartile range around the median (red line), blue whiskers extend to the most extreme data points, red crosses show outliers and red whiskers show the 95% confidence interval calculated for the mean value.

In Part 2 all ESMA configurations were perceived to match good with the dummy head recording. However, it was found that the height layers need to be encoded at a higher gain level as well as a higher elevation angle. It is assumed that the missing difference between the ESMAs with and without height layers is due to this.

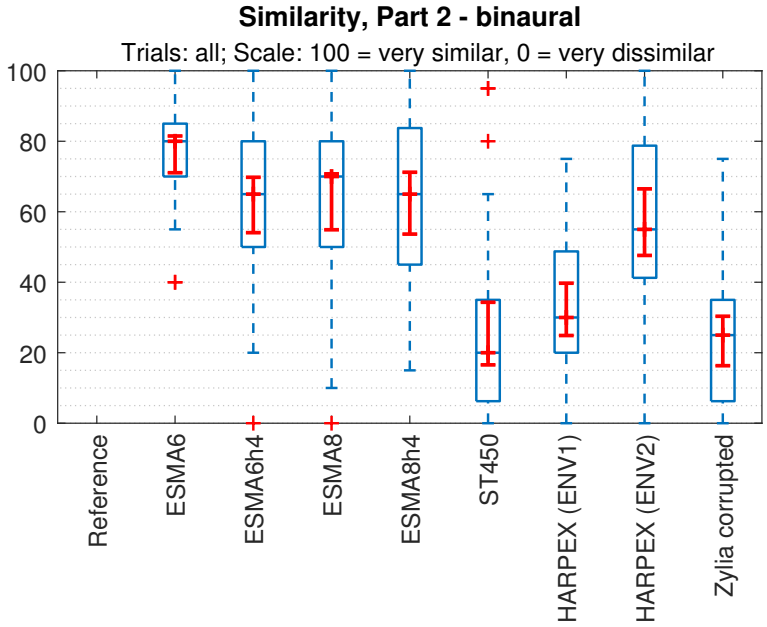


Figure 1.20 – Perceived similarity to the dummy head reference in Part 2. Blue boxes show the inter-quartile range around the median, blue whiskers extend to the most extreme data points, red crosses show outliers and red whiskers show the 95% confidence interval calculated for the mean value.

Since the HARPEX parameter setting with a higher envelopment factor (ENV2) was rated to be very similar to the reference for some listeners it was added as stimulus to the further tests.

A further comparison was performed by analysing the median ranking order of each participant. In figure 1.21 the overall ranking distribution regarding naturalness is shown. Separately for each stimulus, the ranks are plotted on the x-axis and the number of corresponding ratings on the y-axis. The KU100 was hereby rated the most natural recording by 9 out of 12 evaluated listeners. All ESMAs are found mostly in the first half whereas for some listneres the ESMA6 performed even more natural than the dummy head. The majority rated the ST450 at positions 6-9. It is noted again that the Zylia is excluded from the comparative judgement because of the corrupted decoding.

It was reported by participants that the trials without additional speech sample were rather difficult to rate. Consequently stimuli will be selected with higher thoroughness. Another pilot test outcome was that a reference is essential. The large data spread could have probably been reduced by better scale labelling and by limiting binaural playback to only one headphone model.

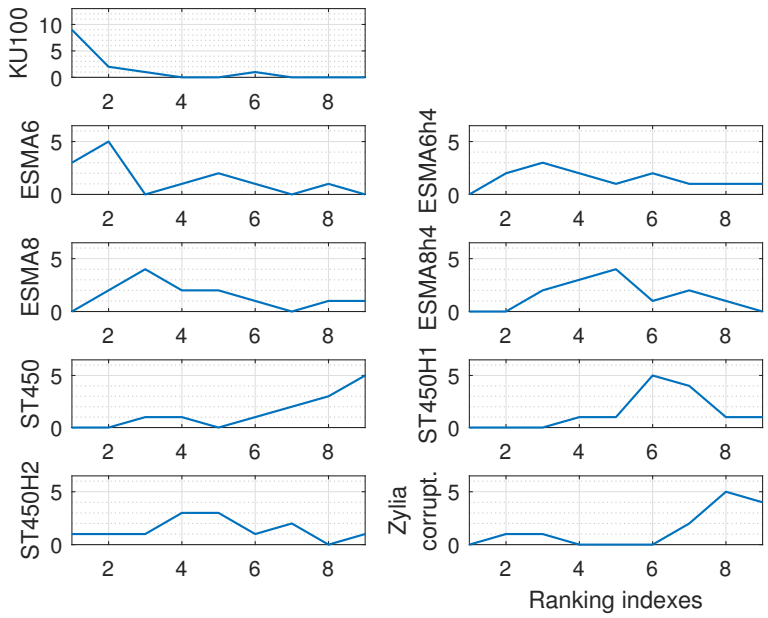


Figure 1.21 – Ranking indexes summed over all subjects for the first part

Based on the findings of this preliminary test, the microphone arrays are examined in detail in chapter 4.

1.3 Recording setup and scenes

The 'plausible scene recording' chapter concludes with a short discussion on the selection of suitable acoustic scenes for the target application. Additionally, an overview of the used recording setup is given.

In general, acoustic scenes can be divided into open-air and indoor scenes, while the second differs mainly in room size and the number of people present. The psychoacoustic hearing experiments carried out at the UWO are dedicated to human perceptions in everyday scenes - especially those of hearing impaired persons. They mainly struggle to understand speech in environments which feature numerous speakers, additional noise or high reverberation levels [SKWD19]. The scenes to be selected should therefore cover a wide variety of places where one could find oneself in everyday life. At the same time, different room sizes and complexities of acoustic scenes should be included. An additional prerequisite is that spoken language should only be present as babble noise, but not in terms of distinct conversations. It turns out that the latter is the main problem for finding small venues and for getting authorisations to record in non public spaces.

Following the results of the conducted listening experiments in chapter 4, the ESMA8h4 and the Zylia microphone are chosen to be part of the scene recording setup - thus every microphone array category is represented. In summary, this includes on the one hand a very natural sounding but expensive and difficult to assemble ESMA8h4, and on the other hand a comparably inexpensive and easy-to-install Zylia, which, however, involves tonal drawbacks due to coincident recording technology and MEMS capsules.

Since there is no need to synchronize the recordings, the two microphone arrays can be recorded independently. A flap impulse within each take should enable a rough alignment of the parallel recordings. For the ESMA8h4, which contains 12 channels, two Zoom F8 field recorders [31] were used, as they offer the possibility of sample wise synchronization. (Unfortunately no field recorder is equipped with a higher XLR channel amount). The Zylia was directly recorded via USB in Reaper. The front of the microphone marked with a red dot was hereby directed towards the first ESMA8h4 channel.

As the ESMA is much more complicated and time-consuming to transport and set up, only three indoor and one outdoor venue were selected to be recorded with both arrays:

— Office
— Restaurant
— Mall
— Playground

In addition to the venues mentioned above, other scenes were recorded using only the Zylia due to simplicity and weather conditions:

— Crossing
— Trainstation
— Trainstation hall
— Public space (tram/bus station)
— Park

All above listed scenes were recorded from 9-11th November 2019 in Graz, Austria. A documentation of all recording spots and detailed scene description including present sound sources can be found in Appendix A.

31. Zoom F8 field recorder, https://www.zoom.co.jp/products/handy-recorder/zoom-f8-multitrack-field-recorder

Chapter 2

Parameter estimation

The estimation of acoustic parameters attempts to gather as much information as necessary to characterize the acoustics of a place or room out of a recording at only one single microphone position. Ideally, a simplistic measurement could serve data acquisition in a time-efficient way without a lot of additional and expensive equipment. There should be no need for permits, like in the case of using a gun to measure the impulse response and the acquisition should moreover not depend on the availability of a power supply. If the type and estimation of the parameters is accomplished with sufficient accuracy, an artificially generated virtual room of relatively low processing effort and natural sounding signals can be successfully reproduced, either to embed augmenting sounds in a given recording or a fully artificial room.

There exist several techniques that enable estimation of room acoustical parameters without any reference in laboratory conditions. Because the noise floor is high in typical real-life ambient scenes, it is difficult to extract parameters without any additional information about the recording. Since in this master thesis the parameter estimation is to be developed as an application for practice, it was decided to use a simple impulse of a wooden start flap to gain information at manageable acquisition. The necessary material is cheap and easy to procure, but yet reasonable and efficient.

Nevertheless balloon pops should be mentioned here as a possible alternative. According to Abel, room impulse responses (RIRs) estimated from balloon pop recordings seem to match loudspeaker measured RIRs after applying processing to expand the frequency bandwidth [ABH$^+$10]. Furthermore Seetharaman and Tarzia [ST12] showed that on average, similar results were obtained with clapping compared to balloon pop measurements - at least in the octave bands of 1-2 kHz in quiet environments. However, according to them, only a SNR of 26.4 dB was reached. Furthermore, their result implies an averaging over several claps.

By recording the start flap, a rough impulse response is acquired. Compared to common methods to measure an impulse response, the SNR is lower and as depicted in figure 2.1 the power is limited in the frequency range. Since the application is aimed at busy recording venues, a low SNR to the environmental noise must be expected. For example, a SNR of 10 dB would only allow the 2-4 kHz octave bands to be included. Therefore, no broadband excitation can be assumed. Furthermore it is supposed that the frequency responses of flaps differ depending on their construction, which is why no equalization is applied.

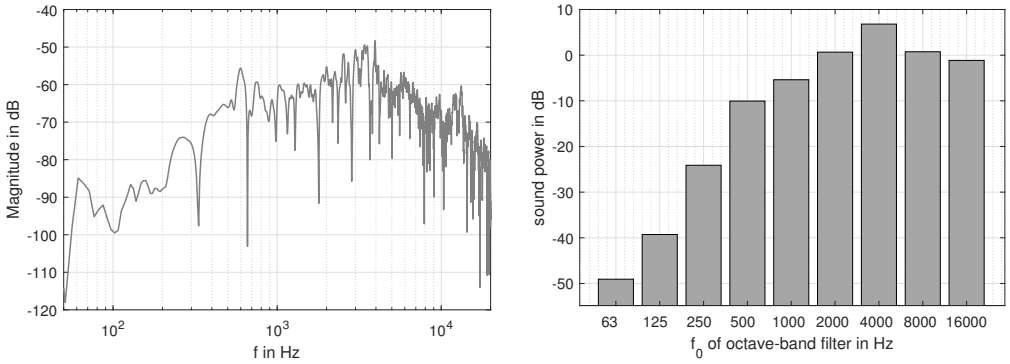


Figure 2.1 – Left: FFT magnitude of a normalised flap impulse, measured in the anechoic measuring laboratory at the IEM. Right: Sound power per octave band of the same flap impulse. (RMS-normalized, no physical normalisation.)

However, it shall be shown how well the parameter estimation works without the need of elaborate equipment and in which regard this approach limits the achievable reproduction accuracy. For the implementation, an Ambisonics signal recorded with the Zylia microphone is used. With a SNR of 69 dB, the microphone lies within a realistic range between sufficient quality and affordable price.

The sound in a room depends strongly on the position of both source and receiver relative to room boundaries and other acoustically relevant surfaces. In order to meet the requirement of a source embedding, that is adapted to the room and thus plausible, it is useful to divide the room into radial sectors around the microphone and to record an flap-impulse in each sector. Number and size of these sectors depend on the structure of the room and its acoustic complexity. Four to eight room sectors are assumed to provide a nice coverage and sufficiently many data points.

During the recording the start flap was held in front of the body at an average height of 1.5 m and directed towards the microphone array. A description of the room including dimensions and an overview of the chosen sectors and positions can be found in Figure 4.3 in section 4.2.1. In the following analysis the impulse response of one detected flap

impulse at a distance of 3.5 m and an azimuthal angle of -18° will be examined exemplarily. In the first graph in figure 2.2 the decay is shown by means of the absolute value of the recorded impulse. For better visibility of the peaks and contours, the absolute square of the first 50 ms is depicted.

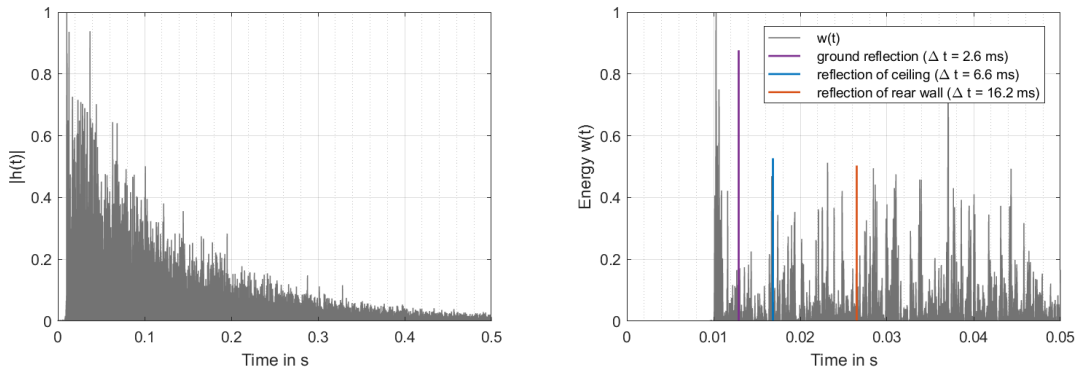


Figure 2.2 – Impulse response $|h(t)|$ obtained through the flap at a distance of 3.5m and an angle of -18° (azimuth) in a calm ambience (left) and a detailed view of the first 50ms of $w(t) = h^2(t)$ including depicted reflections (right). Normalisation was performed regarding the maximal peak.

The peaks arriving before the arrival time of the ground reflection are reflections on the body of the person operating the flap. During the measurement it was observed that care must be taken to operate the flap with the same intensity for each impulse to prevent variations.

As a ground truth to compare with, RIRs were measured in the same room using the exponential sine sweep method and a loudspeaker. To measure the discrete impulse response $h[n]$ a cyclic deconvolution can be performed on the linear system response $y[n]$ and the source signal $s[n]$ containing the exponential sine sweep

$$h[n] = IFFT\{\frac{FFT\{y[n]\}}{FFT\{s[n]\}}\}. \tag{2.1}$$

Conversely to the start-flap impulse response, figure 2.3 shows far more distinct reflections. Moreover the ground reflection is more attenuated in this case, which could be due to differing radiation patterns of the flap versus the Genelec 8020 loudspeaker employed in the measurement.

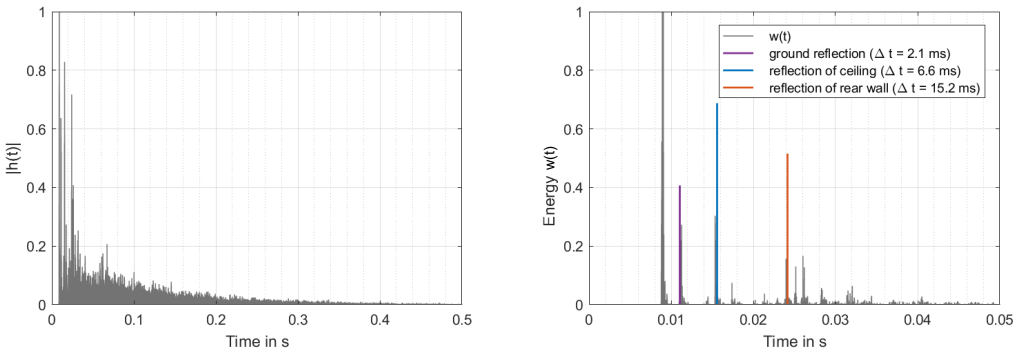


Figure 2.3 – Impulse response $|h(t)|$ obtained through an exponential sine sweep method (left) and a detailed view of the first 50ms of $w(t) = h^2(t)$ including distinct reflections (right). For playback a Genelec 8020 loudspeaker was used, which was located at a distance of 3m, at an angle of -5°(azimuth) and a height of 1.3m. The recording was made in a calm ambience. Normalisation was performed regarding the maximal peak.

The following sections show how much information can still be obtained from such a rough start flap impulse response and which advantages can be drawn from an Ambisonics impulse response in this context. Since no frequency-dependent statements can be made for all octave bands, the relevant information of such a rough impulse response is contained in

— how fast the emitted energy decays,
— the energy ratios between direct sound, early and late reflections and
— the directions from which most energy comes.

2.1 Zeroth-order Ambisonics Room Impulse Response (ARIR)

2.1.1 Reverberation time

The reverberation time T_{60} in room acoustics is the time in which the total sound pressure level drops from 0 dB(SPL) to -60 dB(SPL) after pulse excitation.

Usually, and especially with regard to estimates from recordings of places with a higher environmental noise, a 30 dB decay is measured and extrapolated instead of measuring the 60 dB decay.

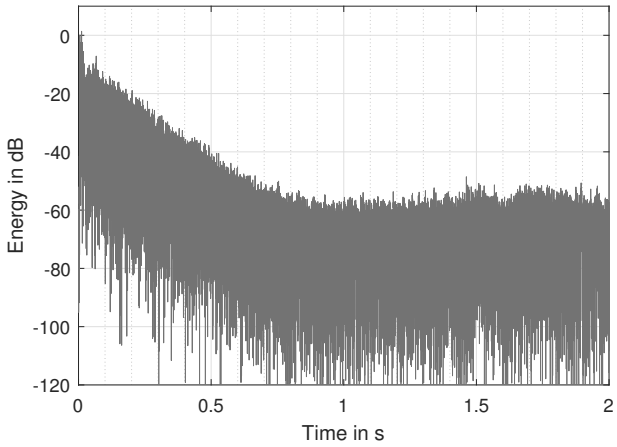


Figure 2.4 – Energy decay in dB of the recorded flap impulse.

For this purpose, a smoothing of the room impulse response is performed via a moving average filter (window length = 50 samples). Subsequently, the EDC (Energy Decay Curve) is calculated via the Schroeder backwards integration

$$EDC(t) \triangleq \int_t^\infty h^2(t)dt. \tag{2.2}$$

To determine the reverberation time from the EDC, however, a finite time has to be chosen as integration limit. This parameter is quite essential, because if the background noise floor is included within the integration interval, the expected exponential slope may be corrupted by a bias leading to a raised linear ramp. It can be determined through various methods, following e.g. Lundeby, Chu or Hirata. For this rough estimation the simple estimation by calculating the RMS value of the noise floor was chosen.

The reverberation time T_{30} is calculated through

$$T_{30} = 2 \cdot (\mathrm{EDC}^{-1}\{-35dB\} - \mathrm{EDC}^{-1}\{-5dB\}), \tag{2.3}$$

for a normalized EDC so that $\mathrm{EDC}(0) = 1$.

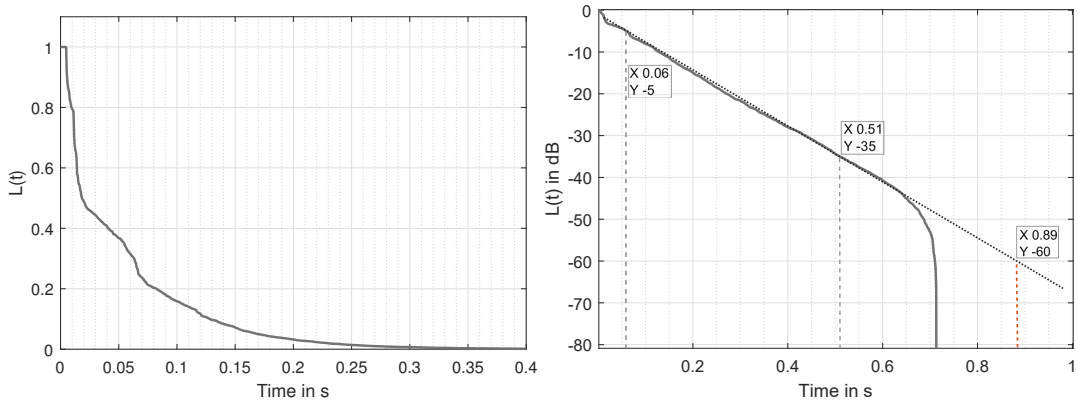


Figure 2.5 – Energy decay curve $L(t)$ (left) and in dB (right) with depicted t's for -5 and -35dB to calculate $T_{30} = 0.89s$.. The black dashed line shows the $-60dB$ limit.

In the following, the calculated T_{30} values are given in table 2.1 for the zeroth-order channel of the Zylia microphone and, as a comparison, also for the NTI measurement microphone.

T_{30}	Zylia ZM-1	NTI
-18°, 3.5m	1.01 s	1.08 s
153°, 1.8m	0.96 s	0.91 s

Table 2.1 – Calculated T_{30} for the Zylia ZM-1 and the NTI measurement microphone on two different flap impulse positions in the office.

2.1.2 Energy ratios

The room impression is essentially affected by the energy ratio of direct sound, early reflections and late reflections. Thus, by determining two ratios, characteristics of a room of predominant and fundamental importance can be considered covered.

Above all and especially with regard to the perception of distance, the most important measure is the direct-to-reverberant (D/R) ratio. It is defined as energy ratio of the direct sound to the diffuse sound

$$D/R = \frac{\int_{0ms}^{T} s^2(t)dt}{\int_{T}^{\infty} s^2(t)dt}, \tag{2.4}$$

by using the zeroth-order Ambisonics channel for $s(t)$ and a time constant $T = 1.5$ms regarding only direct sound.

Regarding the flap an omnidirectional radiation pattern was assumed, so that γ_{src} is set to 1 for all calculations hereinafter. In table 2.1.2, the calculated D/R values are again compared for the Zylia and the NTI measurement microphone.

D/R	Zylia ZM-1	NTI
-18°, 3.5m	-11.9 dB	-12.5 dB
153°, 1.8m	-8 dB	-10.2 dB

Table 2.2 – Calculated D/R for the omnidirectional signal of the Zylia ZM-1 microphone and the NTI measurement microphone on two different flap impulse positions in the office.

A second important energy ratio is the 'Hallmaß', which is generally defined at 1kHz as

$$H = 10log\Big(\frac{E_\infty - E_{50}}{E_{50}}\Big) \quad \text{in dB.} \tag{2.5}$$

However, in this thesis, a modified version not including the direct sound will be used, in order to establish an energy ratio between early and late reverberation

$$H_T = \frac{\int_{Tms}^{50ms} s^2(t)dt}{\int_{50}^{\infty} s^2(t)dt}. \tag{2.6}$$

In table 2.3 estimated values by using the proposed equation 2.1.2 are presented.

H_T	Zylia ZM-1	NTI
-18°, 3.5m	-2.1 dB	-1.5 dB
153°, 1.8m	-1.27 dB	-0.9 dB

Table 2.3 – Calculated adapted 'Hallmaß' H_T for the omnidirectional signal of the Zylia ZM-1 microphone and the NTI measurement microphone on two different flap impulse positions in the office.

2.2 First-order and higher-order ARIR

2.2.1 Direction-of-arrival (DOA) estimation

As mentioned in the introduction of this chapter, directional energy distribution plays an important role in the perception of space, in particular in the early part of the room impulse response. Most current approaches use the image source method, where a shoebox room of suitable proportions is typically modelled. However, without including further acoustic properties specific to the room, such as absorption and scatter coefficients, this offers only a rough approximation.

Furthermore, the method in this thesis is intended to work without prior knowledge about room geometrics. For this reason a DOA estimate is used to generate a fingerprint of the room by detecting direction, strength and time delay of the early and prominent reflections. Usually 1st and 2nd order reflections, as illustrated in figure 2.6, are considered early and the following reflections late, but the splitting point, also called mixing time, is rather arbitrary.

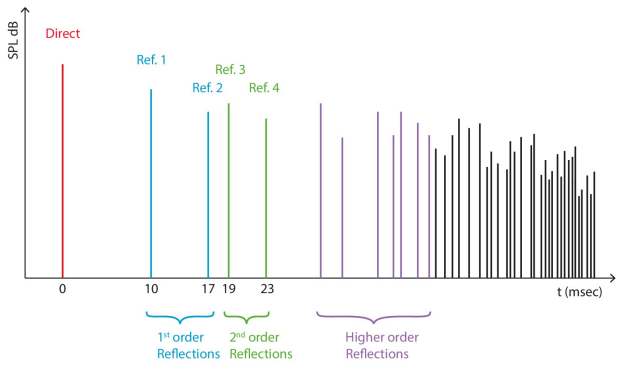


Figure 2.6 – Reflectogram showing arrival times of 1st, 2nd and higher order reflections for an exemplary room impulse response [2].

2. Reflectogram, https://odeon.dk/learn/articles/room-acoustics/

For this purpose, intensity vectors are calculated per time frame. In case of the Zylia ZM-1 Ambisonics signal, the omnidirectional channel is used for the non-directional pressure component. The X, Y and Z channels are interpreted as gradient microphones and represent the velocity part

$$\boldsymbol{r}_{DOA}(t) = W(t) \begin{bmatrix} X(t) \\ Y(t) \\ Z(t) \end{bmatrix}, \quad \boldsymbol{\Theta}_{DOA}(t) = \frac{\boldsymbol{r}_{DOA}(t)}{||\boldsymbol{r}_{DOA}(t)||}. \tag{2.7}$$

as suggested in [MZZ20].

Care must be taken, that the microphone capsules cannot be placed in the same place, which means that spatial aliasing is created at higher frequencies. Since directional information is mainly transmitted over 200 Hz, a bandpass filtered signal (200Hz - 4kHz) is used for the DOA estimation. In order not to over-represent quiet signal components, filtering should be applied before normalization.

The implementation is based on an earlier seminar paper on sharpening directions via DOA estimation [EFW18]. The 1st-order Ambisonics signal is buffered in time frames of 2 ms length for each of which one intensity vector is calculated.

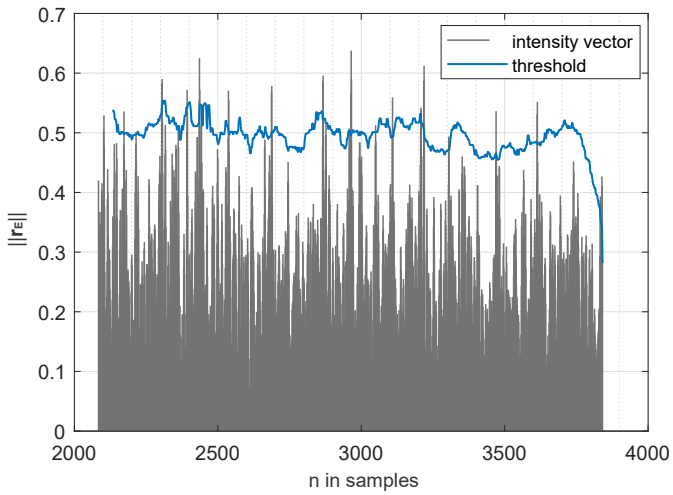


Figure 2.7 – Threshold calculated via a median filter (order = 100). The noisegate, raising the threshold level, was set to 0.25.

Depending on the desired number of detected events, a threshold (figure 2.7) is applied on the length of the intensity vector over time. It is defined via the median filter order and an additional noise gate. To determine the direction of the flap impulse, a high threshold would be used.

In order to obtain a desired number k of reflections, the threshold is defined by using the mean free path

$$L = \frac{4 \cdot V}{S} \approx \frac{4}{T_{30}} \quad \text{and} \quad k \approx \frac{t_{ER} \cdot c}{L}, \tag{2.8}$$

with t_{ER} denoting the time duration in s for the early reflections. (It is referred to section 2.2.2 regarding estimation of the room volume.)

Thus direction, time delay and strength of the discrete early reflections within t_{ER} can be determined. Figure 2.8 shows second and higher-order early reflections selected by the defined threshold plotted on a sphere.

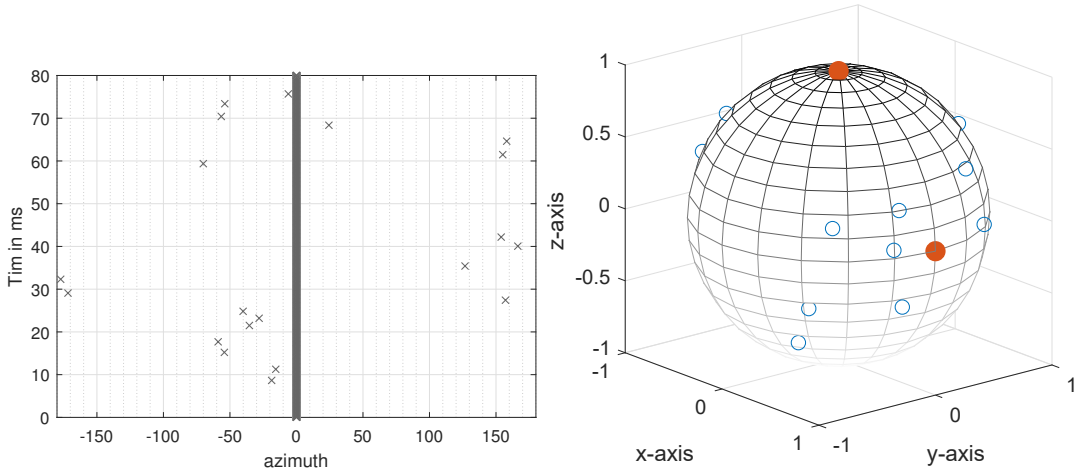


Figure 2.8 – Crosses show detected early reflections in a trace plot for the azimuthal angle (left). Blue circles show the spherical coordinates of those reflections between 3 and 80ms. The DOA estimation is based on the impulse response of the examined flap at an azimuthal angle of $-18°$. Orange spots mark locations of coordinate axes x and z.

2.2.2 Directionally sampled ARIR

It has proven useful to estimate first and higher orders of early reflections separately, since the former can still be distinctly identified. Hereby it is assumed that an accurate reproduction of the first-order reflections (typically 6 in a shoebox shaped room) plays a more important role. For this reason, beams are used to sample the ARIR only in the respective directions. Subsequently time delays $\Delta t_{1..6}$, directions $\phi_{1..6}$ (azimut) and $\theta_{1..6}$ (elevation) and corresponding energies can be estimated.

Starting from the microphone, one first-order beams is directed towards the ground

($\theta_1 = -45°$) and one towards the ceiling ($\theta_2 = 60°$) in the azimuthal direction of the detected flap (via DOA estimation). Using maximum peak detection, the squared impulse response of the resulting spatial section is examined for peaks representing ground and ceiling reflections. The horizontal plane is sampled in 10 degree steps by a third-order beam (see figure 2.9). Simultaneously, the four highest wall reflection peaks are detected.

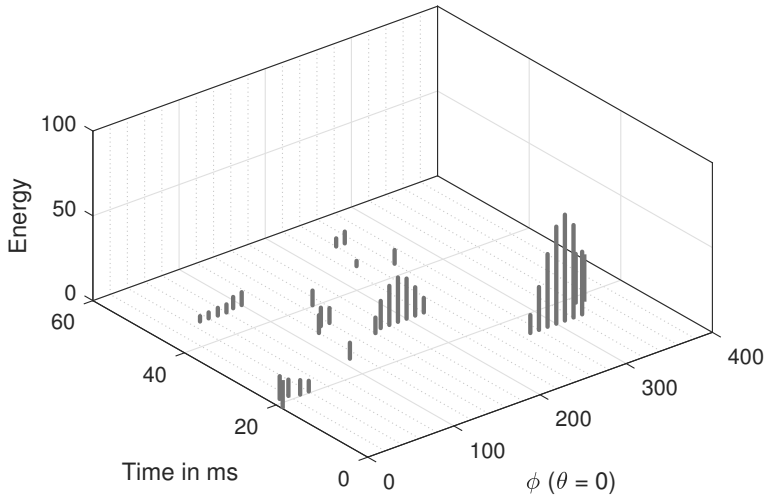


Figure 2.9 – Energy of wall reflections determined via a 3rd-order beam, ϕ in 10 degree steps.

Furthermore, with knowledge of Δt_1 and the assumption that the height of the operated flap h_{flap} is always the same, the distance d_{flap} can be calculated by using the geometrical relations

$$d_{gr} = 2 \cdot \sqrt{\left(\frac{d_{flap}}{2}\right)^2 + h_{flap}^2} \quad \text{and} \quad d_{flap} = d_{gr} - \Delta t \frac{c}{Fs} \quad \text{in m.} \tag{2.9}$$

An alternative method would be to record the flap impulses on a fixed radius, if detection of Δt_1 is impossible due to a high environmental noise.

Additionally, also the room volume can be derived from d_{flap} and the D/R ratio measured for the flap impulse position. A generally valid and distance-dependent D/R ratio is to be derived as

$$D/R = \frac{|p_{dir}|^2}{E\{|p_{diff}|^2\}}\gamma_{src} = \frac{\frac{1}{(4\pi r)^2}}{\frac{1}{A\pi}}\gamma_{src} = 0.0032\frac{V}{T}\frac{1}{r^2}\gamma_{src}, \tag{2.10}$$

with γ_{src} denoting the directivity factor of the source and $A = 0.161\frac{V}{T}$ [GM04].

In the implementation the zeroth-order of the Ambisonics signal was hereby used to full fill the need of an omnidirectional directivity pattern of the receiver. The directivity factor γ_m for the microphone will be considered to be 1, correspondingly.

This equation is also known to be used in order to calculate the critical distance r_H, where the energy of direct and diffuse sound are the same,

$$D/R = 1 = 0.0032\frac{V}{T}\frac{1}{r_H^2}\gamma_{src}, \quad r_H = 0.057\sqrt{\frac{V}{T}}\sqrt{\gamma_{src}}. \tag{2.11}$$

For the estimated parameters of the respective flap impulse, the volume of the room can now be roughly calculated via

$$V = \frac{1}{0.0032} \cdot D/R \cdot T_{30} \cdot d_{flap}^2 \cdot \frac{1}{\gamma_{src}}. \tag{2.12}$$

And finally, by using geometric relationships, all further room dimensions (height, width and length) could be estimated, taking into account directions and time intervals as well as the relative positions of microphone and flap. However, as this did not bring any advantage for the developed embedding method, no further investigations were carried out in this regard.

Chapter 3

Embedding

This chapter integrates the outcomes of the two previous chapters. On the one hand, we gained an ambient scene recording which was optimized to sound as plausible as possible. On the other hand, it was described which parameters can be estimated from a rough impulse response based on a recording with a HOA microphone. The collected data from the parameter estimation shall now be used to embed an arbitrary sound source into the ambient recording.

The input parameters are merely distance d, azimuth ϕ and elevation angle θ as well as the source width. Depending on the resulting position a corresponding room sector is selected, for which the parameter estimation is thus performed. This gives the following additional values: T_{30}, D/R ratio, H_T, direction, time delay and gain of the early reflections, the approximate room volume and an average number of early reflections. Due to the limited frequency range of a flap impulse, no frequency dependencies are included.

In the following, the embedding approach implemented in MATLAB is presented. The structure of the chapter follows the composition of the resulting convolution-based sound: direct sound, early reflections and late reflections.

3.1 Direct sound

The embedding of direct sound provides an important contribution to the perception of source direction and distance. Since the embedding is primarily designed for reproduction in a loudspeaker dome and the focus in this work is on static sources, directional panning of the direct sound at ϕ and θ is to be done via VBAP [ZF19a]. For an extension to dynamic sources moving at defined trajectories an encoding in Ambisonics would be the method of choice.

To adjust the level depending on the system and the sound source in general, a normalization factor g_N is set for a reference distance of 1 m. The level of the direct source, depending on the embedding distance d, is thus controlled by the relation

$$g = \frac{1}{d} \cdot g_N. \tag{3.1}$$

All further relative levels are derived from the system gain weighted source. The ratio of is values and target values of D/R and H_T are subsequently applied as gain factors for the early and late reflection signal part. Appropriate energy relation inherently models the apparent source width, which according to Lee decreases almost linearly per doubling of the source-receiver distance [Lee13].

3.2 Early reflections

As introduced in section 2.2.1, a recording of a flap in characteristic room sectors around the microphone forms the basis for the embedding of early reflections. Automatic segmentation of the flap impulse recording and directional detection selects the corresponding room sector according to the embedding directions ϕ and θ.

Determined directions, time delays and gains from the directional sampling of the ARIR (first-order reflections) and the DOA estimation (higher order early reflections) are used to encode the source signal in Ambisonics. For the latter, a time interval t_{ER} starting from the mean delay of the wall reflections until 80 ms was chosen [Lin14].

Since ground and ceiling reflection make a significant contribution to localization, the azimuthal direction is taken from the embedding position. Regarding the ground reflection also time delay and attenuation are calculated via geometric relations of the embedding distance d. The elevation angle for encoding of the ground reflection α_{gr} is thus given by the relationship

$$\alpha_{gr} = \arctan\Big(\frac{h_{flap}}{\frac{d}{2}}\Big). \tag{3.2}$$

A ratio of the time delay depending on the source-receiver distance is shown in figure 3.1.

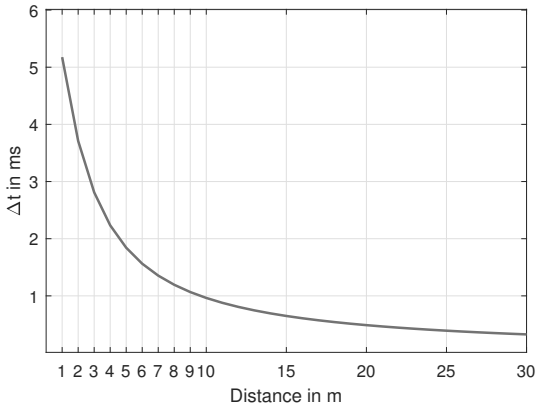


Figure 3.1 – Ratio of the source - receiver distance and the time difference Δt_1 of arrival times between direct sound and ground reflection.

In order to simulate the higher absorption of wall surfaces above 1 kHz [1] [Vor07], a high-shelf filtered version (biquad filter; Fc = 5000Hz, Q = 0.6, G = -5dB) is used as input signal.

To avoid phasing or comb filter effects, a decorrelation as described in [Blo19] is additionally applied. In the method implemented there, see GUI in figure 3.2, an arbitrary number of impulse responses with random group delay are generated. Here, each reflection is hence convolved with one response from a pool of those impulse responses.

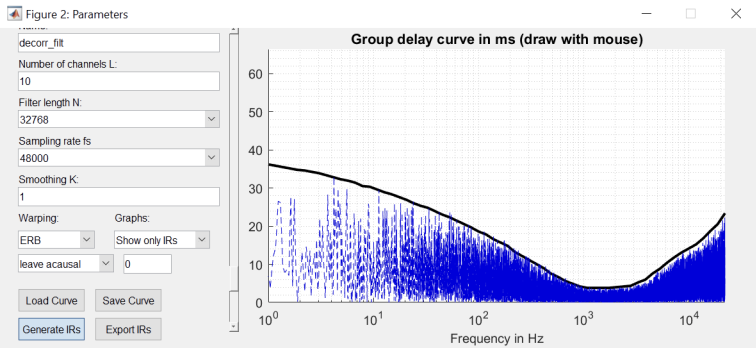


Figure 3.2 – Matlab tool for impulse response generation. [Blo19]

1. Absorbtion coefficient database, https://www.ptb.de/cms/ptb/fachabteilungen/abt1/fb-16/ag-163/absorption-coefficient-database.html

Since it has been observed that later reflections are often unnaturally strong, the gains of the higher-order early reflections are weighted with a decaying Gauss function. Additionally, in order to achieve a consistent ratio within the early reflections, the maximum peak of the second-order reflections was limited to the first-order maximum.

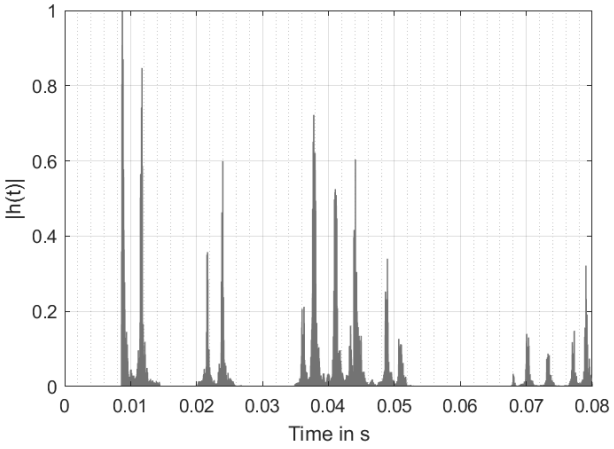


Figure 3.3 – Resulting simulated BRIR (right channel) regarding direct sound and early reflections.

3.3 Late reflections

In contrast to a classical convolution reverb, feedback delay networks offer a comparatively effective and flexible possibility to model diffuse fields in a physically plausible way. It was first introduced by Stautner and Puckette and offers a high echo density.

As already shown in [Blo17], such a network can be used to create an Ambisonic reverb. If a sufficiently high order is used, the often suggested modulation in the feedback path in order to avoid a typical blaring sound can be omitted. The approach was implemented by S. Grill [Gri17] and Daniel Rudrich in an open source VST-plugin [2], and is the basis for the implementation in MATLAB in this master thesis.

For generating the late reflections via the algorithm shown in figure 3.4, a mono sound file encoded in 7th-order Ambisonics at the predefined embedding direction is fed into a 64-channel FDN. (Comment: This was corrected after the listening test was carried out. Currently the output of the early reflections, encoded in Ambisonics, is used as input.) To simulate a realistic fade-in of the late reflections, two structurally identical FDNs are

2. Open-source code, FDN plugin, https://git.iem.at/audioplugins/IEMPluginSuite/-/blob/master/resources/FeedbackDelayNetwork.h

calculated simultaneously, with the second FDN adjusted to a a shorter reverberation time of 0.1 s and then subtracted from the original signal.

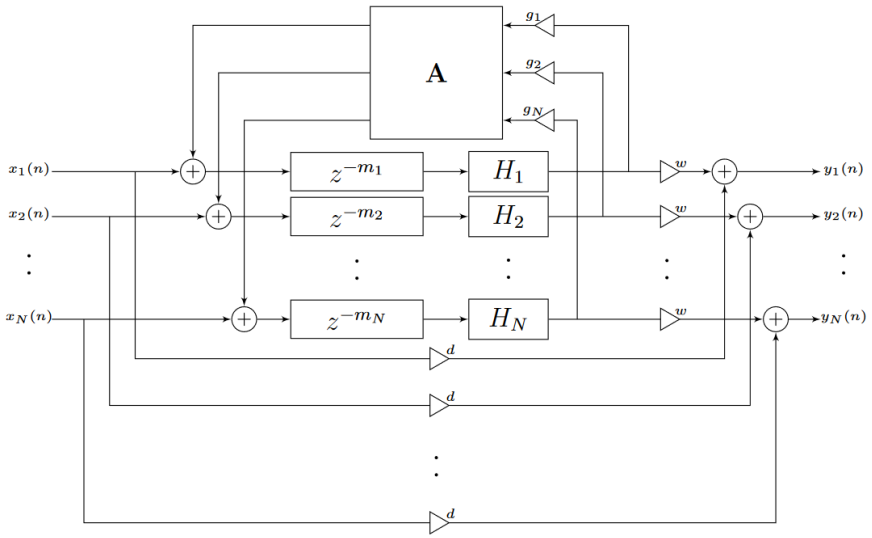


Figure 3.4 – Flow graph of the implemented Feedback Delay Network [Gri17].

Input parameters are in this case the reverberation time T_{30} and a factor R for the room size, ranging from 5-30, which could be determined through a dependency on the previously calculated volume.

The individual delay lengths $m_{1..64}$ are defined by using a sequence of selected prime numbers in order to avoid periodicities

```
delayLen = primes(roomSize/10:roomSize)./10*1000 *Fs; %samples
```

Each channel in the forward path is filtered by a low-shelf and a high-shelf filter, implemented as IIR filter. The cut-off frequency was set for both filters to $F_c = 1$ kHz. Since the filter gains cannot be controlled by a frequency-dependent reverberation time in this case, $g_{low} = 0$ dB and $g_{high} = -70$ dB.

For an accurate control of the reverberation time [SH17] the corresponding linear gain is converted depending on the respective channel delay via the relation

```
function gain_conv = channelGainConversion(delayLen, Fs, gain_input)
    len = delayLen./Fs;
    gain_conv = gain_input.^len;
end
```

The overall gain $g = 10^{\frac{-60}{20 \cdot T_{30}}}$ describes the attenuation per second and it is adopted to the individual sample delay in each channel of the FDN.

In order to enable control of the reverberation time via this function, the feedback matrix A has to be lossless. An efficient solution also for a real time application is offered by an orthogonal, normalized Hadamard matrix, which was proposed by Rochesso for calculations in real-time [Roc97]. Due to its structure, this matrix distributes the energy very diffusely to the channels.

```
A = hadamard(64)/8;
```

Since only the wet signal will be used, the gain factor d (dry) is set to '0' and w (wet) contrary to '1'.

Finally, because the resulting reverberation would still be too dominant for frequencies above 1 kHz, an additional shelving filter (G = -8 dB, Q = 0.404, F_c = 3000) is applied.

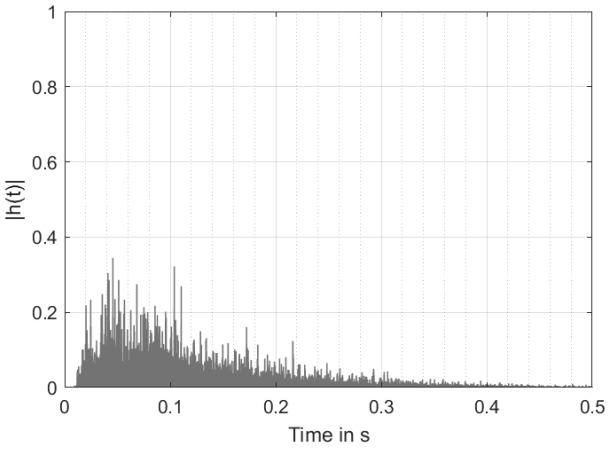


Figure 3.5 – Resulting simulated BRIR (right channel) for the late reflections obtained via a FDN.

3.4 Overall result and outlook

In combination with a HOA microphone, also a coarse impulsive excitation signal seems to provide useful parameters. It has also been demonstrated that DOA-based convolution and an ambisonic FDN reverb are valuable tools for source embedding with the estimated parameters.

Rumsey stated [Rum02] that for a plausible spatial impression it is not necessary to model every reflection accurately. Correct modelling of energy ratios play a much more

important role in this regard. Especially strength ratios within the early reflections and the corresponding time differences are significant. In this point the presented method could be enhanced by adding further constraints. Nevertheless, by using a DOA estimation, a trade-off between accuracy and plausibility is made, according to my observations.

As a drawback of the sketched approach, the flap as excitation signal appears to overestimate the energy of early reflections, whereas the energy of the direct sound is underestimated. It is quite likely that this is caused by a non-omnidirectional radiation, contrary to the assumption made. In this respect, balloon pops would offer an advantage in terms of omnidirectional sound propagation.

Concerning the embedding of the late reflections via an artificial FDN reverb, it is noticeable that trebles must be damped relatively strongly to achieve a natural decay. With an additional high-shelving filter, see description from above, the targeted reverberation time can be reproduced. To close the gap to general validity, definition of a relation between the FDN parameter *room size* and the room volume V is still pending and seems to be a rather complex problem following [SH19]. Schlecht also suggests replacing the scalar feedback matrix with a delay feedback matrix, which would introduce a different delay for each feedback matrix entry and refers to the echo density profiles in [AH06].

The proposed method could still benefit from parameter estimation in the frequency domain. In the case of a broadband excitation signal, the energy ratios as well as the frequency decay could be determined in octave bands. A more complex but ideal solution would be to use battery-operated spherical speakers for measurements with exponential sine sweeps.

A listening experiment presented in section 4.3 could give further insight. The results will be examined with regard to plausibility and naturalness in acoustic perception.

Chapter 4

Evaluation

Finding a proper test design supporting meaningful results strong enough to test hypotheses may often turn out to be difficult. First of all, unambiguous and preferably uni-dimensional attributes play a crucial role in the selectivity of the test design and hereby support a desirable reproducible indication of differences.

As described by Rumsey [Rum02] one can distinguish generally into two test design concepts. On the one hand the classical psychophysics method including very simple stimuli like noise bursts e.g. for the purpose of studying how human brain reacts on sound and to highlight very specific differences. Comparative judgements on the other hand are more suitable to comparatively evaluate multiple 'products', where subjects are used as quality meters.

The term *ecological validity* describes *"the extent to which an experimental situation matches the real-world context and circumstances it is supposed to represent"* [Rum02, p.654]. To ensure ecological validity, the audio material used should correspond to the target application in a reproducible format. The reproduction setup of the listening test should at least come as close as possible to the perceptual characteristics of the 91-channel loudspeaker dome.

4.1 Statistics

Statistics were planned following an overview of statistical methods for audio experiments given in [JH15].

For the statistical analysis significance tests will be carried out, which are either parametric or non parametric depending on the data distribution. If the experiment data are not normally distributed the Wilcoxon pair test and the Friedman-Test have to be

computed instead of a pairwise t-Test and ANOVA test.

$h = 1$...rejection of H_0 at the significance niveau $\alpha = 0.05$ (no normal distribution)
$h = 0$...failure to reject H_0 at the significance niveau $\alpha = 0.05$ (normal distribution)

Data	**parametric** normal distributed	**non parametric** not normal distributed
2 independent samples	t-Test	Mann-Whitney-U Test
2 dependent samples	pairwise t-Test	Wilcoxon pair test
more than 2 independent samples	ANOVA	Kruskal-Wallis-Test
more than 2 dependent samples	ANOVA (rep. measures)	Friedman-Test

Table 4.1 – Overview of statistical tests for parametric and non parametric tests.

If necessary further evidence can be obtained by investigating effect sizes. In comparison to significances, these are independent of the number of test subjects and the cumulated alpha-error. The commonly used value describing effect sizes for normal distributed data is *Cohen's d* [Coh88]. Values can be interpreted as follows:

d< 0.2...small
d≈ 0.5...medium
d> 0.8...big.

Cliff's δ enables an effect size calculation also of ordinal data sets, which additionally do not have to be necessarily normal distributed [Jam10]. Values of *Cliff's* δ can be interpreted as follows, according to [RKCS06]:

$|\delta| < 0.1$...neglectable
$|\delta| \approx 0.1$...small
$|\delta| \approx 0.3$...medium
$|\delta| > 0.5$...big.

Hypothesising

For all conducted tests the null hypothesis H_0 is that no difference by means of plausibility, spatial impression or distance is perceptible between all microphone array stimuli. In contrast the alternative hypothesis H_A states that there are significant differences.

Following the results of the pilot test it is supposed that one of the ESMA arrays will be rated the best at least regarding naturalness, and that all higher-order Ambisonics signals will overrule the first-order microphone.

Paradigma

A MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) design including rating based scales was decided to be used.

Generally the MUSHRA Test is structured into parts and herein trials which are used to examine repeatedly on a specific rating question. A labeled reference is presented in each trial together with test samples, a hidden reference and an anchor, whereas the purpose of the latter is to calibrate the scale.

Because all stimuli are presented in a parallel manner fewer participants are required to obtain significant results compared to other statistic test. To gain more repeatable and consistent results with even fewer participants the test is intended to be executed by expert listeners, which also have a better internalization of the scale. Results are assumed to be valid also for consumers since in any case preferences are expected to be the same for experts and naive listeners. It was at least shown that the resulting rank order is the same [SB13], even though Rumsey showed that trained listeners weigh spatial artefacts stronger contrary to naives who focuse primary on timbral characteristics [FRK05]. Subsequently about 20 experienced listeners should participate in the test for significant results. To show location parameters such as median and mean values as well as confidence intervals, at least 30 samples per stimulus should be obtained. Moreover, to keep the effect of the alpha-error cumulation low, 5-7 stimuli should be presented maximum at each page.

A standalone MUSHRA App, implemented by Rudrich at the IEM, was chosen to be used for conducting the listening experiment. It works together with an OSC-message receivable audio host, e.g. Reaper, which enables real-time audio processing, for example to use headtracking. The MUSHRA App is easy to configure via a .JSON file and generates Matlab readable output files which is the reason for which it was used for setting up the listening experiment. [1]

4.2 Plausibility of microphone arrays

The first part of the evaluation chapter targets transferring the acoustic reality with one of the proposed microphone arrays to a spherical loudspeaker setup in laboratory conditions. To detect which array matches the desired plausibility criteria best a listening experiment was conducted.

1. MUSHRA App, Rudrich (IEM internal repo), https://git.iem.at/rudrich/mushra

Parameters that circumscribe the desired plausibility need to be established to enable well-posed tasks in the test. Many attempts have been made to define and categorise attributes for various reproduction applications, including music live recordings and hyper-real VR content. To mention some: Toole 1980s [Too85], for the Spat at IRCAM (differ between soure-related, room-related) [JPJW93] and Rumsey (developed scene based approach [Rum02]). A vocabulary for describing the perception of virtual acoustic environments in a comparative assessment is proposed in the *Spatial Audio Quality Inventory* (SAQI) [Lin15]. The manual provides a table of categorised perceptual qualities together with a circumscription and scale labels. Qualities of interest for this listening experiment are:

— Distance
— Externalisation
— Level of reverberation
— Envelopment
— Clarity
— Speech intelligibility
— Naturalness
— Presence
— Degree-of-Liking

Another classification was done by Letowski [Let89] within the multilevel auditory assessment language (MURAL) and is shown in in figure 4.2.

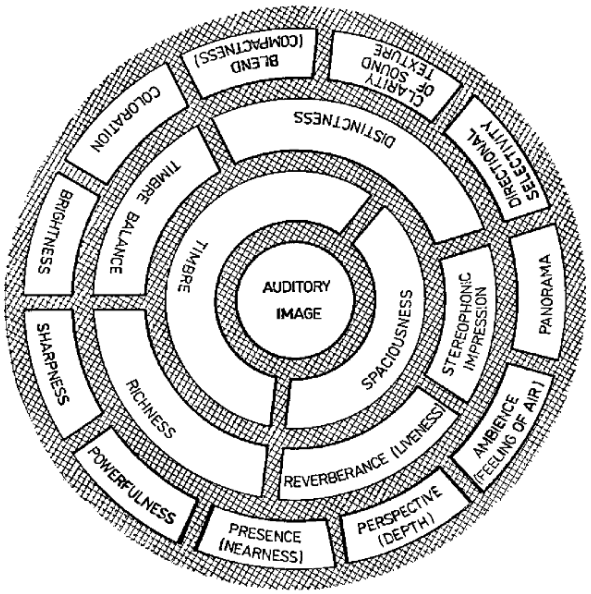


Figure 4.1 – MURAL classification [Let89].

The outer ring MURAL classification terms *clarity of sound texture*, *directional selectivity* and *panorama* complete the above introduced plausibility attributes for microphone array recording. To conclude, those terms could be summarised in three groups: naturalness, spatial impression and distance.

4.2.1 Stimuli preparation

In order to obtain a natural, but nevertheless sample-exactly reproducible environmental scene, an electroacoustically simulated scene was created. Hereby, an office scene was chosen because the bandwidth of the existing sound sources and the accessibility and availability of the same office was also given for recordings during the quiet night time.

Figure 4.2 – Picture of the microphone array setup in the stimuli recording session.

Typical office sounds were recorded in a studio and mixed to nine 1-minute audio files:

- Source 1: typing, chair sounds, writing, cutting
- Source 2: door, keys, bell
- Source 3: steps, bin sounds, typing, table knocking
- Source 4: copier, office table sounds, office chair sound
- Source 5: page turning, human sounds (coughing, laughing,..), typing
- Source 6: speaking, office table sounds, phone ringing
- Source 7: low office sounds, babble
- Source 8: office table tapping, office chair

For the audio scene simulation, the composed sound sources were spatially distributed to eight Genelec 8020 loudspeakers of various heights (depicted by red loudspeaker symbols 1-8 in figure 4.3). All of them were positioned to face a wall in order to record diffused rather than direct sound. One further loudspeaker, which was moved to four positions of various angles and distances served as distinct source of direct sound (blue loudspeakers 1-4).

Figure 4.3 – Floor plan including schematic furniture of the stimuli recording location. ESMA8h4 ring and dummy head are depicted including correct orientations. The four employed loudspeaker positions of RIR measurements are illustrated in blue. Five clapping positions are marked by black crosses and define corresponding room sectors backed light blue. Red loudspeakers depict sound sources which were used to simulate the virtual office scene.

Since the stimuli recording session was also used for the data acquisition of the parameter estimation, the room was divided into 6 sectors (light blue areas). In every room sector a flap impulse at positions 1-6 (marked with black crosses) was recorded.

Regarding the microphone array recording setup the KU100 was positioned at the normed ear height of a sitting person (1.2 m) typical for the recorded office scene. It was rotationally aligned with the frontal direction of the Ambisonics microphones. The ESMA-3D ring was placed slightly beyond at a height of 1.15 m. Channel 1 of each ESMA configuration started in a relative angle of 200° to the frontal dummy head direction. To compare several horizontal microphone spacings for the ESMA6h4 and the ESMA8h4, both were recorded at three different ring diameters (0.8 m, 1.24 m and 1.36 m). The Zylia was positioned directly above the dummy head (1.5 m height). On top the SoundField ST450 MK II was mounted via a overhead microphone tripod at a height of 2 m. To monitor the sound level a calibrated NTI measurement microphone was added to the recording setup placed at a height of 1 m right below the dummy head.

A protocol of the recording session including an equipment list and an overview of the recorded audio files can be found in Appendix B. To sum up the relevant recordings: for all six possible ESMA-3D configurations the simulated office scene was recorded separately and together with the speech sample played via loudspeaker at positions 1-4 (depicted in blue in figure 4.3).

Figure 4.4 – Recording session with loudspeakers simulating an office ambience scene.

4.2.2 Test design

Although the targeted playback setup is a 91-channel loudspeaker dome, the listening experiment was decided to be conducted binaurally. Advantages of the binaural test

setup are that a dummy head can be used as a reference of human hearing, the fine directional resolution and that the experiment is available online for the participants. A loudspeaker-based test on a smaller setup will be conducted then to cross-validate the binaural results.

Stimuli used in the listening experiment were on the one hand two different excerpts of the simulated office scene (Atmo 1, Atmo 2), and on the other hand several snippets of a male speaker sample at three frontal positions and one lateral position.

The listening experiment was divided into three parts with a different focus, whereas in every trial a reference recording was presented. The scale from 0-100 could be rated in 20 subdivisions due to a step size of 5. The parts were presented in the same sequence from 1 to 3, instead the trials were randomized. The tasks were given in German, but will be translated herein into English.

Naturalness, Part 1

The first part consists of four trials: Atmo 1, Atmo 2, Speaker at left (3 m) and Speaker at front (3 m).

> *Please rate the naturalness of the binaural stimuli from 'identical' to 'very different' - the reference represents the basic truth of naturalness.*
>
> *The following attributes should be included in the rating process:*
> — *Clarity of the acoustic scene (are all details heard in the reference recording also audible in the stimuli?)*
> — *Directional selectivity*
> — *Presence (feeling spatially integrated)*
> — *Liking of the recording*
> — *Externalisation*

Spatial Room impression, Part 2

The second part consists of three trials: Speaker at front (2 m), Speaker at left (3 m), Speaker at front (4.5 m).

It was decided to present only stimuli with an additional speaker, because results of the pilot test showed that room impression rating is otherwise difficult.

> *Upon entering a room, subconsciously properties that characterize the room acoustically are collected and combined to form an overall spatial impression.*
> *In the following, evaluate the similarity of the stimuli to the reference in terms of the spatial impression - please note that you will hear a different the same excerpt for the reference and the stimuli.*

Parameters such as reverberation time, speech intelligibility and compactness of the sources can be included.

In this case the direction of the sources should NOT be evaluated! Please make sure that 'identcal' in this part is in the middle of the scale - the difference to the reference can be evaluated in the direction of either 'dry' or 'reverberant'.

Distance, Part 3

The third part consists of one trial: Speaker at front at all available distances (2m , 3m , 4.5m).

Sort all stimuli according to the perceived distance of the speaker (in front) from 'closest' to 'furthest'. The reference where the speaker is positioned at a distance of 3 m marks the middle of the scale.

The reference divided the scale into two halves which gives the participants a better understanding of the scale and reproduces more reproducible results. The label of the boundaries includes that the listeners should use the whole scale for sorting.

In this part the ESMA8h4 was excluded in order to reduce the amount of stimuli. It was assumed that ESMA6h4 and ESMA8h4 would resemble each other.

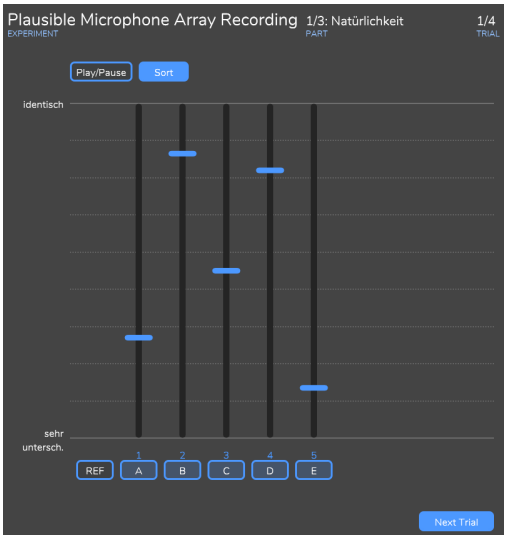


Figure 4.5 – Screenshot of the MUSHRA GUI and an exemplaric rating for the first part.

4.2.3 Binaural experiment

Before the stimuli files were rendered for the listening experiment, the samples were processed in MATLAB and Reaper.

Binaural decoding was done by convolving virtual loudspeaker signals with measured KU 100 dummy head HRTFs contained in the SADIE II database [2] [CAK18]. The ESMA channels were decoded via a direct convolution, whereas the height channels were encoded -10dB lower than the horizontal channels. Accordingly the Ambisonics microphones were decoded via a Sampling Ambisonics Decoder (SAD).

To obtain the spherical harmonics and the max-rE weights MATLAB routines by Archontis Plotis were used [3], based on [ZF12]

$$\boldsymbol{x} = \sqrt{\frac{4\pi}{L}} \boldsymbol{Y}_N^T \cdot diag\{\boldsymbol{a}_N\} \cdot \boldsymbol{y}_N(\theta_s) \cdot s. \tag{4.1}$$

For the decoding of the ST450 FOA recording six virtual loudspeaker positions (azimuth and elevation in degree) were used

$$\boldsymbol{\theta}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\theta}_2 = \begin{pmatrix} 180 \\ 0 \end{pmatrix}, \boldsymbol{\theta}_3 = \begin{pmatrix} 90 \\ -45 \end{pmatrix}, \boldsymbol{\theta}_4 = \begin{pmatrix} 90 \\ 45 \end{pmatrix}, \boldsymbol{\theta}_5 = \begin{pmatrix} 270 \\ -45 \end{pmatrix} \text{ and } \boldsymbol{\theta}_6 = \begin{pmatrix} 270 \\ 45 \end{pmatrix}.$$

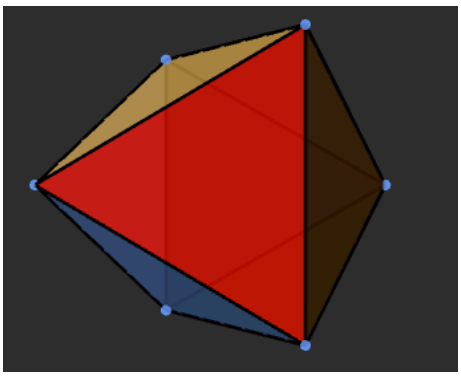


Figure 4.6 – Visualisation of the virtual loudspeaker positions for the ST450 decoding.

The zeroth-order channel of the Ambisonics signal was hereby weighted with $\frac{1}{3}$ to convert from the FuMa to N3D normalisation.

2. SADIE II database, https://www.york.ac.uk/sadie-project/database.html
3. polarch, https://github.com/polarch/Higher-Order-Ambisonics

The HOA microphones were encoded to a 20 loudspeaker t-design layout. Since the Zylia is already max-rE weighted within the Zylia plugin, it just needs to be converted from SN3D to N3D normalisation via an additional gain depending on the respective order.

$$\boldsymbol{g} = diag\{[\sqrt{2n+1}]_n\} \tag{4.2}$$

In order to gain an audio file that better resembles the timbre of the dummy head an equalization function was applied to the Zylia ZM-1 and the SoundField ST450 signal in MATLAB. For this the left and right ear signals were summed and a transfer function H calculated between the third-band smoothed long-term spectrum of the dummy head and the binaural decoded signals of the Ambisonics microphones. Each time frame of the short-time fourier transformed signals (Hann window, window size = 4096, overlap = 1024) was then multiplied with H.
For all other binaural decoded recordings better results could be achieved by equalizing the samples within Reaper by hand. A lot of effort was also put in adjusting the levels of all stimuli manually to enable a good comparative basis.

The Ambisonics signals of the SoundField ST450, the HARPEX-upmix and the Zylia ZM-1 were warped downwards according to their respective vertical position within the recording setup by using he SceneRotator of the IEM Plugin Suite. The plugin was also used to align all stimuli rotationally with the frontal speech loudspeaker as reference point.

Since no head-tracking was used for the static binaural playback during the experiment participants could also take part from home due to the online availability. The only restriction was that they have an Beyerdynamic DT770 headphone at hand. A detailed explanation of software settings was provided together with all needed data via a download link. The listeners were asked to sit in a quiet background environment and to set the system dependent sound level to an appropriate volume.

Post-screening

At stage 1 the ability of the subjects to give consistent and reliable answers is checked. This can be judged from the mean square error term in the one-way ANOVA results [SM09]. The square root was taken in order to comply with the original scale.

Regarding the introduced threshold, Subjects 1 and 7 in Part 1 and Subject 6, 9 and 16 in Part 2 could be excluded. However, this would not change results of the conducted pairwise significance test no subjected is excluded a this stage. It should be noted that the relatively high RMS error is caused by the fact that the data points of each stimulus involve different spatial positions instead of unmodified repetitions.

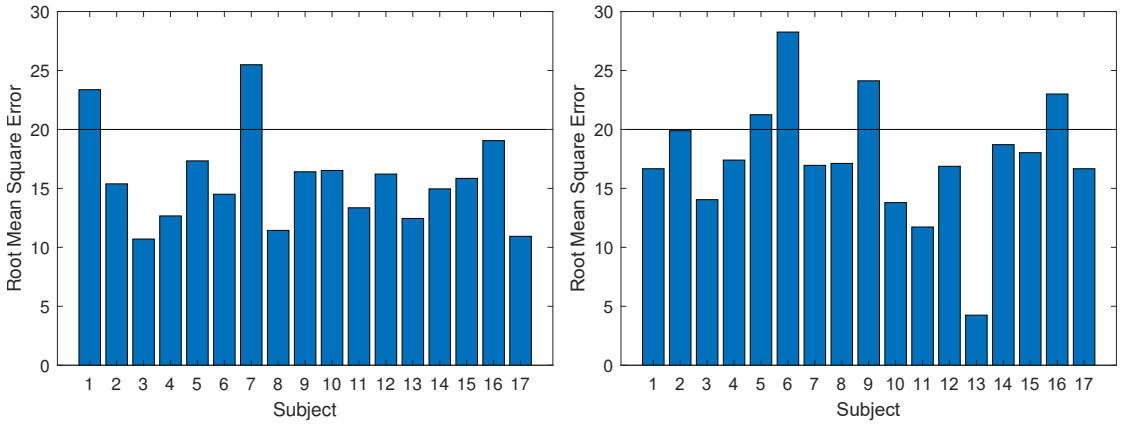


Figure 4.7 – Scaled RMS error for all 17 participants, computed on the evaluations of perceived naturalness (Part 1, left) and spatial impression (Part 2, right). A threshold of acceptability was set at 20%.

At a second evaluation stage subjects which do not rate the hidden reference correctly in Part 2 (*Spatial impression*) for 2 out of 3 trials will be excluded from evaluation. This is the case for Subjects 5, 10 and 17.

Data

In the binaural experiment, conducted in October 2019, 18 subjects (14 male, 4 female) took part which were all sound engineering students and employees of the IEM (Institute of Electronic Music and Acoustics). Similar to the pilot test participants were between 22 and 58 years old, with an average of 28 years. The average duration was 47 min including individual breaks.

Following the results from the Lilliefors and Jarque-Bera-test, H_0 is not rejected for Part 1 (*Naturalness*) (see figure C.1 in the Appendix). In Part 2 (*Spatial impression*) only the hidden reference was detected to show another distribution than the normal distribution. Since the hidden reference will not be compared to the other stimuli, data of both parts can be processed with statistical test allowed for parametric and normal distributed samples.

Due to the experiment design of a MUSHRA test the acquired samples are dependent - so following the overview table 4.1 a *t-Test* will be used for a pairwise test.

The Lilliefors-test for Part 3 (*Distance*) shows a belonging to the not-normal-distributed family for three stimuli, which can be seen in table C.3 and was additionally examined via QQ-plots. Therefore the Wilcoxon pair-test will be used for the evaluation of this part.

Since not all data are from the normal distribution family, the median is given instead of the mean value in all subsequent representations. (For normal distributions and interval-scaled data, the two location parameters correspond if the sample size is large enough).

Results

The results should show whether subjects perceive significant differences between the microphone arrays or not. To be able to give an answer the probability p is defined, that is the probability that the result of an experiment was found by chance compared to a critical value α. We define that if $p < \alpha = 0.05$ the null hypothesis that no difference can be perceived is to be rejected.

Overall analysis (Naturalness)

There is no significant difference between ESMA6h4 and ESMA8h4 regarding *Naturalness* in Part 1 (table 4.2). This is confirmed by a small value of Cliff's $\delta = 0.096$. All other arrays could be differentiated by the listeners following the *paired t-Test*.

	ST450	Zylia	ESMAh6	ESMAh8
ST450 Harpex	0.009	7.958e-5	1.143e-12	9.411e-14
ST450		5.178e-14	4.068e-18	2.963e-20
Zylia			0.013	0.003
ESMAh6				0.325

Table 4.2 – Table showing the Holm-Bonferroni corrected p-values for Part 1 (Naturalness).

In figure 4.8 both ESMAs outperform the coincident arrays, whereas the Zylia was rated to be more natural than the ST450 HARPEX up-mix. Equally to the the pilot test, the SoundField ST450 was ranked to be most different from the binaural reference. It could be observed that 14 participants rated one of the ESMAs as the most similar array regarding naturalness (the Zylia or the HARPEX was top rated by the other three).

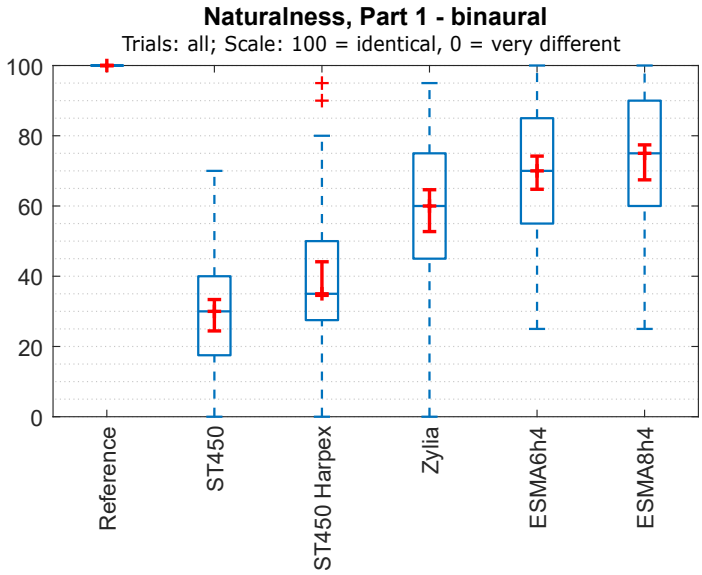


Figure 4.8 – Perceived naturalness in Part 1 of the binaural test. Blue boxes show the inter-quartile range around the median, blue whiskers extend to the most extreme data points, red crosses show outliers and red whiskers show the 95% confidence interval.

Particular scenarios

When analysing and comparing the trials separately in figure 4.9 (whereas the two ambience trials are combined) it seems that evaluation is more ambiguous for the Ambience trials. Interestingly the 'Ambience' and the 'Speaker front' trial resemble each other more. Especially the Zylia, but also the ST450, is rated differently in case of the 'Speaker left' trial. For this trial the ST450 and the ST450 HARPEX differ significantly from each other.

There is also a significant difference perceived between the two ESMAs for the 'Speaker front' trial. This could be related to the orientation depending results of the ESMAs, whether a sound source is located on-axis with one of the microphones or not. Like depicted in figure 1.18 in section 4.2.1 for the ESMA6h4, the position of the speaker at front is in between two microphones, whereas the speaker at left is aligned on-axis. In case of the ESMA8h4 the frontal position is shifted almost on-axis with one microphone. Contrary to the ESMA6h4 the position of 'Speaker left' is not aligned with a microphone of the ESMA8h4.
This leads to the conclusion that a sound source is perceived more naturally when it is on axis with one of the ESMA channels.

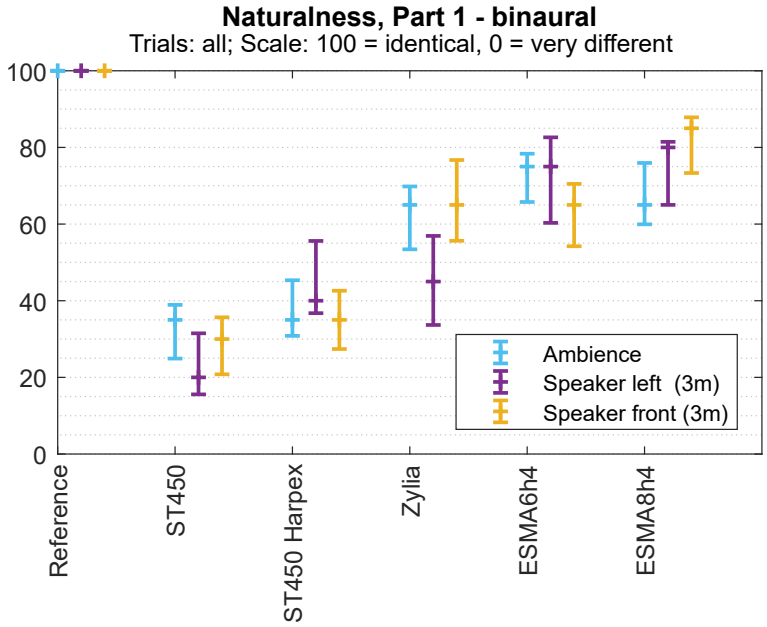


Figure 4.9 – Comparison of presented trials of Part 1 (Naturalness). Whiskers show 95% confidence intervals around medians.

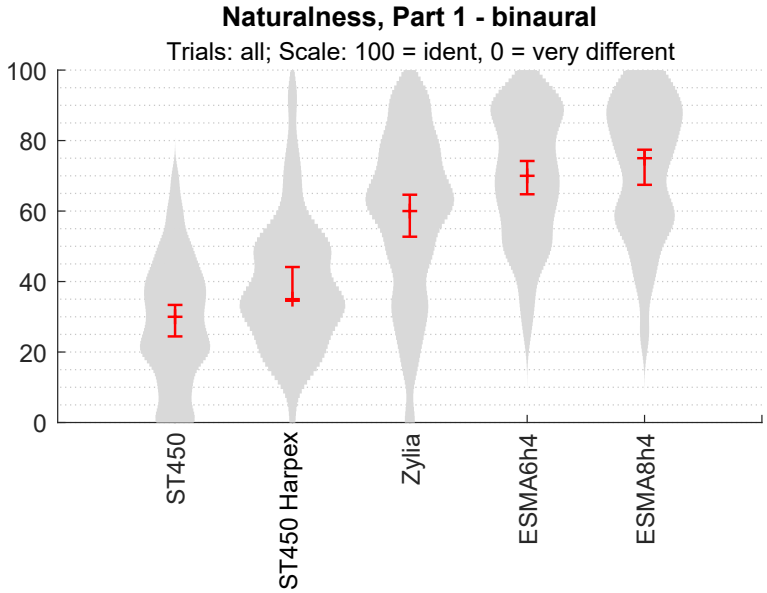


Figure 4.10 – Globally normed violin plot of listeners evaluation of Part 1 (*Naturalness*). All trials and participants were included. Red whiskers show 95% confidence interval.

A violin plot of all trials of Part 1 (*Naturalness*) displayed in figure 4.10 shows a spread of answers over the whole scale in particular for the Zylia. Ratings for the ST450 tend to cumulate at the lower and both ESMAs at the upper scale boundary.

Overall analysis (Spatial Impression)

For Part 2 fewer significant differences are evaluated if all trials are included, as it can be seen in table 4.3. The HARPEX up-mix was perceived to differ from the Zylia and the ESMA8h4, as well as the ESMA6h4 from ESMA8h4.

	HARPEX (ENV2)	Zylia	ESMA6h4	ESMA8h4
ST450	0.350	0.733	0.871	0.467
ST450 Harpex		0.031	0.896	0.017
Zylia			0.896	0.758
ESMA6h4				0.011

Table 4.3 – Table showing the Holm-Bonferroni corrected p-values for Part 2 (*Spatial impression*). Subject 6 is excluded.

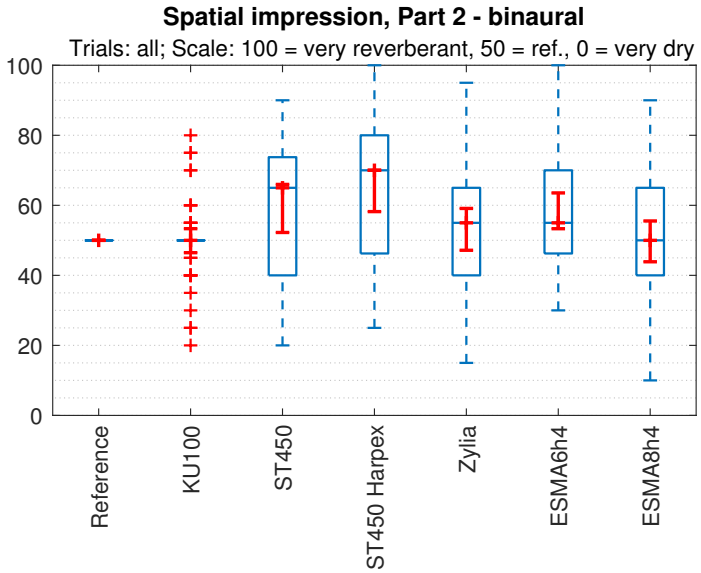


Figure 4.11 – Perceived spatial impression in Part 2 of the binaural test. Blue boxes show the inter-quartile range around the median, blue whiskers extend to the most extreme data points, red crosses show outliers and red whiskers show the 95% confidence interval.

Except for the median of the ESMA8h4, which matches with the reference, every array was rated to be more reverberant than the dummy-head reference (figure 4.11). All FOA based stimuli were rated to be more reverberant than the other arrays (it should be recalled that the ST450 was positioned vertically the highest during the recording). The ESMA8h4 was rated the significantly drier than the ESMA6h4 and tends to be perceived the driest array in general.

The rating of the hidden reference (KU 100) in figure 4.11 offers the possibility to exclude listeners from this part of the experiment (the outliers were mainly produced by 3 participants).

Particular scenarios

Since the perceived spatial impression of the hidden reference is rated similarly to the dummy head reference for all speaker positions, the evaluation should generally be trial independent. However, differences between the trials shouldn't be neglected (see figure 4.12). For both ESMAs the speaker position at front (2 m) produces significantly drier results. The significant high difference for this trial between the ESMA6h4 and the ESMA8h4 can again be explained by the varying relative position of the sound source to the frontal microphone pair. That those differences get smaller when the source is further away, like in case of the 'Speaker front (4.5 m)' trial is reasonable though.

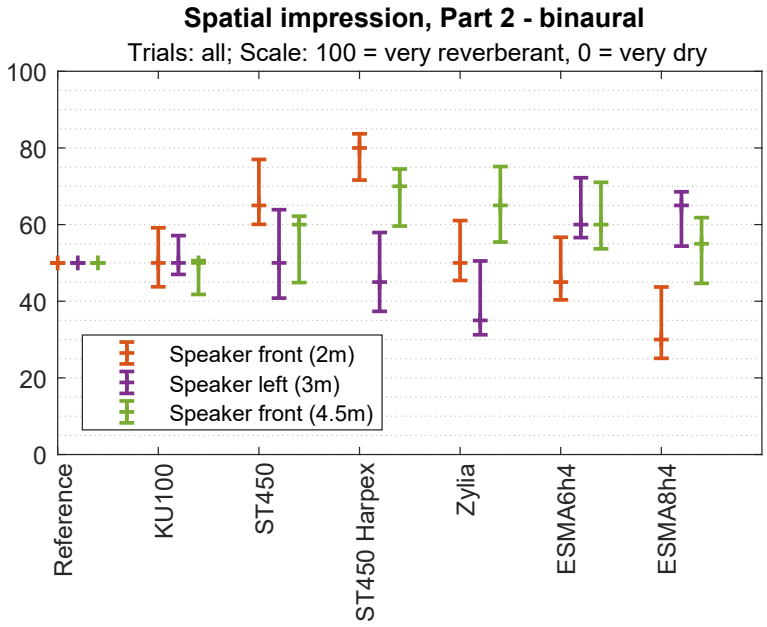


Figure 4.12 – Comparison of presented trials of Part 2 (*Spatial impression*). Whiskers show 95% confidence intervals around medians.

What is surprising is that for both ST450 based stimuli the speaker distance of 2 m is rated to be more reverberant than the 4.5 m distance in comparison to the reference. The speaker position at left tends to be perceived drier than the other two speaker positions for the Ambisonics stimuli, but especially for the higher-order HARPEX and Zylia. The most stable result over all trials (apart from the hidden reference) seems to be achieved by the ESMA6h4.

Whether those variations should be taken into account or not is unclear. It could be argued that participants recognised the hidden-reference KU100 stimuli and simply matched it with the reference. Therefore those trial variations could be still related to various sample snippets out of the voice recording, which could be responsible for a different room exaggeration depending on the present voice pitch and articulation. Another factor is the scale, which can be interpreted differently for each trial.

Auditory distance mapping

Due to the amount of stimuli the impact of the Bonferroni-Holm correction is very high and results in just a few significant differences in Part 3 (*Distance*). A table showing the significances for Part 3 can be found in table C.4 in the Appendix. However, the intention was to have one page where all distances (2m, 3m, 4.5m) of all stimuli can be judged comparatively against each other.

First of all, the ratings in case of the KU 100 reference seem to prove that distance dependent rating is possible. Over all, the reproduced distances by the Zylia microphone are rated very consistently with the dummy-head reference they even seem to be overdone. The median range of the ESMA6h4 sample matches quite good, although it seems that no difference between 3 and 4.5 m is perceivable in comparison the the Zylia stimuli.

All distances of the ST450 and the ST40 Harpex are rated more distant than the reference (figure 4.13). The averaged spacing between those two arrays complies with Part 2 (*Spatial impression*), where the ST450 Harpex is also rated a slightly more reverberant. It is outstanding that ordering is performed consistently wrong for both, with the 4.5m rated equally distant as the 2 m sample. However, this is so far in line with the evaluations of the spatial impression (the position at 4.5 m was evaluated significantly drier than at 2 m) - if one assumes that the reverberation is only a part of the distance impression. Furthermore, the large spread shows that the test persons were very uncertain. Thus, the perceived distance at 4.5 m would be reduced by the difference in the reverberation.

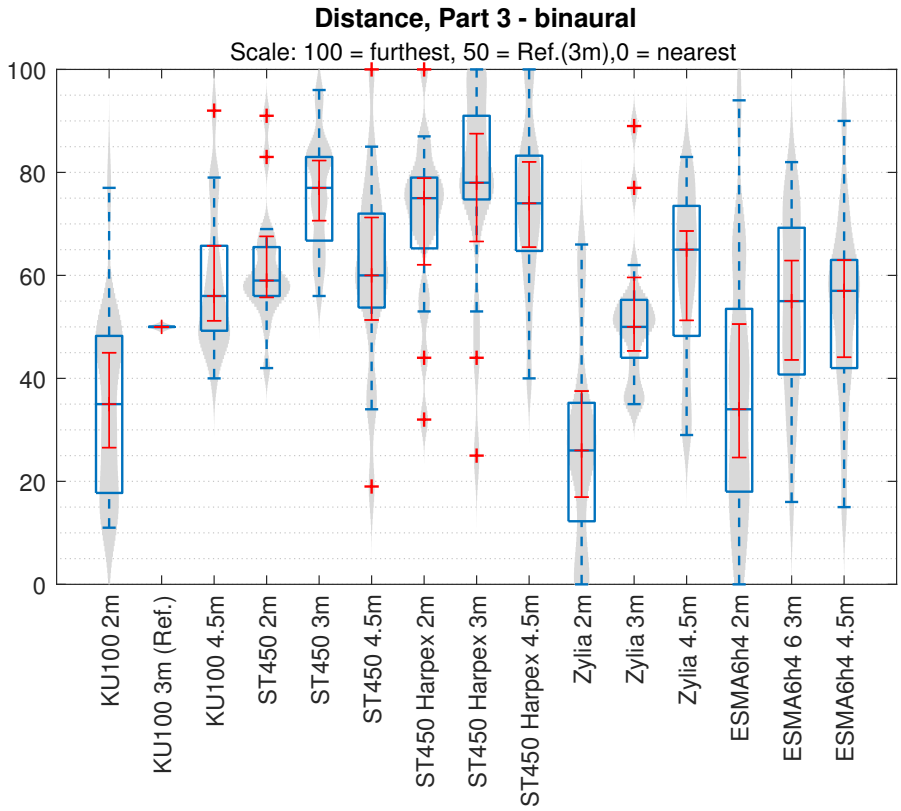


Figure 4.13 – Perceived distances in Part 3 of the binaural test. Blue boxes show the inter-quartile range around the median, blue whiskers extend to the most extreme data points, red crosses show outliers and red whiskers show the 95% confidence interval. The gray shading shows the distribution as violin plot normalized over all stimuli.

In a closer inspection of the individual results it was examined that two participants failed in ordering the dummy head recordings correctly depending on the distance as well as five (out of 17) the ESMA6h4 distances. Quite a few listeners also rated 3 and 4.5 m within one array the same. Interestingly some rated the closest recording, mainly of the ESMA6h4 but also of the dummy head and the Zylia microphone, as the furthermost, leading to a very spread distribution. This could be due to lateral geometric reflection paths emerging for the distant sound and hereby conflicting the auditory distance cues.

It should be noted that some listeners didn't use the whole scale in Part 3 as they were asked in the instructions. However, a normalisation of the data was not regarded to be reasonable.

4.2.4 Loudspeaker-based experiment

In a second experiment the above findings regarding playback via headphones shall now be verified. Ideally, the same tendencies should be visible, even if the decoding can only be done on fewer (now real) loudspeakers. It was conducted in December 2019 including a similar base of expert listeners.

A loudspeaker array with 16 Genelec 8020 was set up within an anechoic measuring room at the IEM to perform a comparative listening test. The loudspeakers were positioned to match all ESMA6h4 and ESMA8h4 horizontal angles. Four height layer loudspeakers were positioned at 30° elevation in between the ESMA6h4 and ESMA8h4 optimal azimuthal position, in order to spare one further loudspeaker.

In the experiment setup all ESMA files were routed directly. The Ambisonics files were decoded via the AllRAD IEM plugin [Rud19], whereas virtual loudspeakers were inserted at both poles with the gain set to 0 as depicted in figure 4.14.

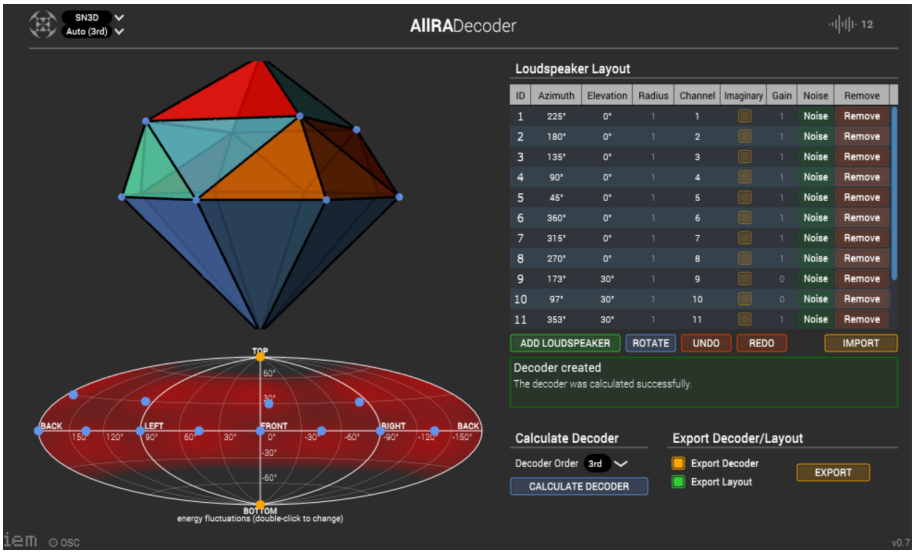


Figure 4.14 – GUI of the AllRAD decoder plugin and settings used in the listening experiment.

At the central position a chair and the control unit was placed, before the ideal position of the head was explained to the participants.

In general, the aim was to make the listening test as comparable and similar as possible. One difference was that this time small head movements were allowed. Additionally, since the binaural reference could not be played via the loudspeaker setup the listening experiment was conducted without reference.

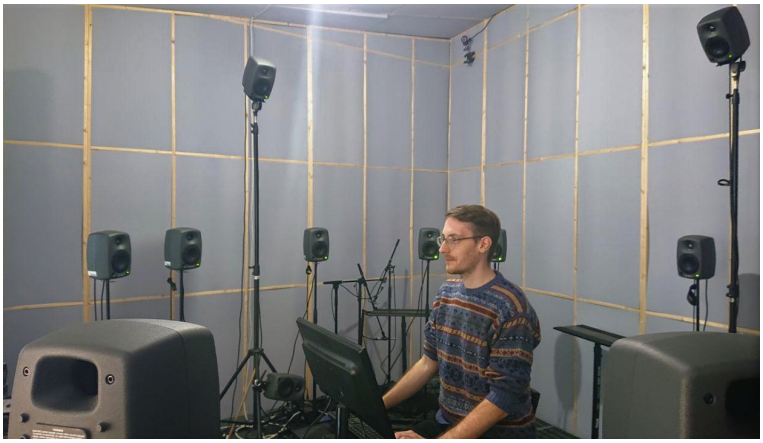


Figure 4.15 – Loudspeaker array with participant.

Naturalness, Part 1

The label was changed from 'identical' and 'very different' to 'very natural' and 'unnatural'.

Spatial Room impression, Part 2

In this part the two categories (reverberant and dry) were negelcted and the instruction was changed to:

> *Sort the stimuli according to the reverberation of the room from 'most reverberant' to 'driest'. Please use the scale and rate the stimulus with the most reverberant room impression with 100 and the driest with 0.*

Distance, Part 3

In Part 3 the Zylia sample at 3 m replaced the dummy head reference. Since a direct comparison between ESMA6h4 and ESMA8h4 seems desirable after analysis of the binaural experiment, all ESMA8h4 stimuli were included and only one ST450 stimulus was retained at 3m.

Post-screening

For Part 1 (*Naturalness*) the intra-listener consistency was much better compared to the binaural experiment. Contrary, evaluation of Part 2 (*Spatial impression*) seemed to be more difficult (see figure 4.16).

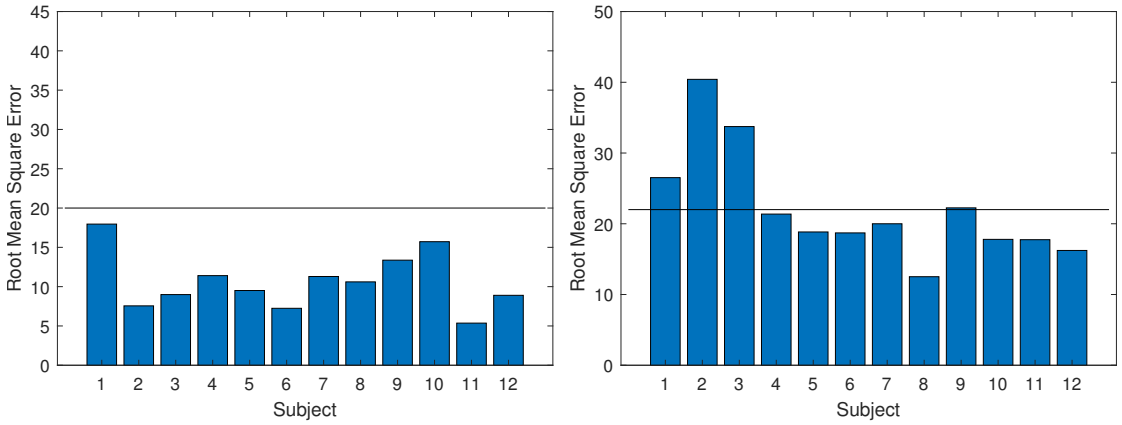


Figure 4.16 – Scaled RMS error for all 17 participants, computed on the evaluations of perceived naturalness (Part 1, left) and spatial impression (Part 2, right). The threshold was set equally to the binaural experiment at 20 and 22.

A closer examination of individual answers of Subject 2 and 3 contradicted a possible exclusion from the evaluation of Part 2 (*Spatial impression*).

Data

In the loudspeaker-based experiment 12 experienced listeners took part, whereas four of them also participated in the binaural experiment. The duration of the experiment was 35 minutes on average.

Following the Jarque-Bera normal distribution test, the results differ from the binaural test. In Part 1 (*Naturalness*) the ST450 Harpex, the Zylia and the ESMA6h4 were not normal distributed. In Part 2 (*Spatial impression*) normal distribution is not applicable for the ST450 and the ST450 HARPEX and in Part 3 *Distance* for the ST450 Harpex (3m).

Results

Results of the loudspeaker based experiment generally show the same tendencies as the binaural test.

Overall analysis (Naturalness)

Pairwise t-tested probabilities in table 4.4 indicate an even higher significant difference. Additionally, data spread seems to be less.

	ST450 Harpex	Zylia	ESMA6h4	ESMA8h4
ST450	0.86086	5.0811e-07	5.0811e-07	1.5863e-08
ST450 Harpex		1.3287e-06	1.7879e-08	1.5863e-08
Zylia			5.0811e-07	9.3998e-07
ESMA6h4				0.86086

Table 4.4 – Table showing the Holm-Bonferroni corrected p-values for Part 1 (*Naturalness*).

For the loudspeaker based test the median and inter-quartile ranges of the ST450 and the HARPEX up-mix as well as of the ESMAs become almost identical.

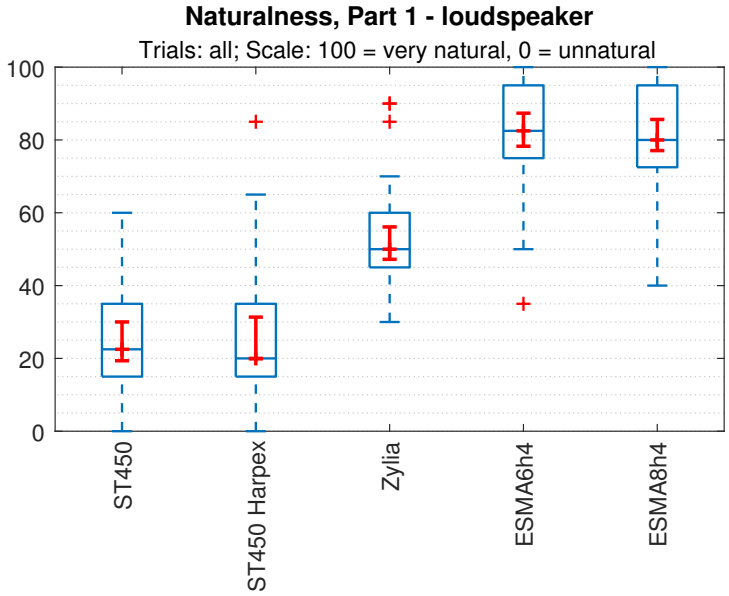


Figure 4.17 – Perceived naturalness in Part 1 of the loudspeaker-based test. Blue boxes show the inter-quartile range around the median, blue whiskers extend to the most extreme data points, red crosses show outliers and red whiskers show the 95% confidence interval.

Particular scenarios

Since no reference as in the binaural experiment was presented, the scale was used now from 0 to 100. However, since fewer subjects participated, the confidence intervals are also larger.

Results for both ESMAs and the Zylia are very stable over all trials. However the 'Ambience' trials seem to be perceived more unnatural in case of the ST450 based stimuli compared to the ones containing a speech sample.

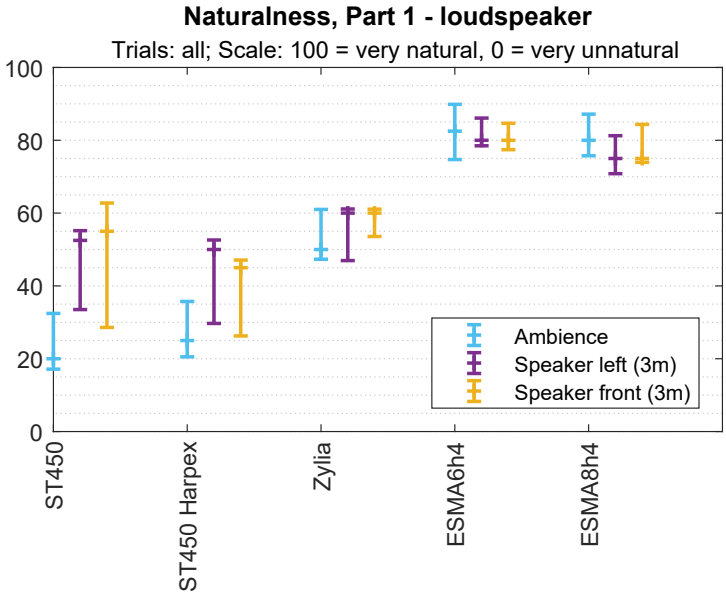


Figure 4.18 – Comparison of presented trials of Part 2 (*Spatial compression*). Whiskers show 95% confidence intervals around medians.

Overall analysis (Spatial impression)

For part 2 only statements in relation to the binaural experiment due to the absence of a reference can be made.

The only significant different stimulus seems to be the ESMA8h4 (except for the Zylia microphone). Between the ESMA6h4 and the ESMA8h4 a *Cliff's* δ of $\delta = 0.3$ is calculated, indicating a medium effect.

	ST450 Harpex	Zylia	ESMA6h4	ESMA8h4
ST450	1	0.136	0.145	0.011
ST450 Harpex		0.291	0.270	0.029
Zylia			1	0.145
ESMA6h4				0.045

Table 4.5 – Table showing the Holm-Bonferroni corrected p-values for Part 2 (spatial compression)

As expected and shown in figure 4.19, a rating of spatial impression without a reference produces very spread data. The same tendency as in the binaural experiment can be observed, that the ESMA8h4 is perceived to be drier.

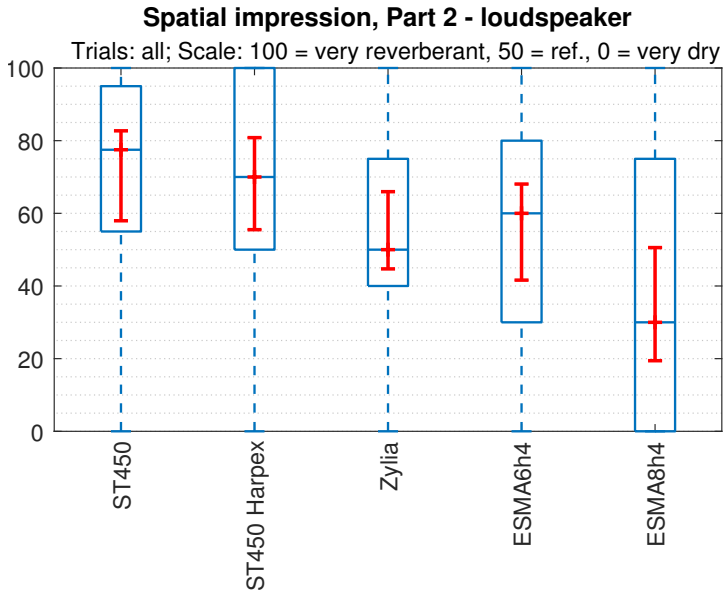


Figure 4.19 – Perceived spatial impression in Part 2 of the binaural test. Blue boxes show the inter-quartile range around the median, blue whiskers extend to the most extreme data points, red crosses show outliers and red whiskers show the 95% confidence interval.

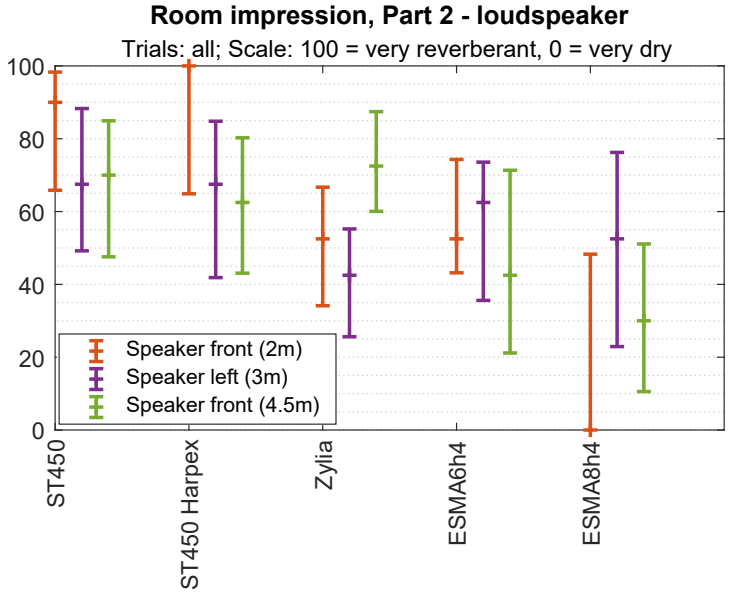


Figure 4.20 – Comparison of presented trials of Part 2 (*Spatial compression*). Whiskers show 95% confidence intervals around medians.

Particular scenarios

Even when analyzing the individual trials, there is a great similarity in the tendencies of the binaural trial for all stimuli. In particular, the room impression for the Zylia microphone was evaluated exactly the same even without reference.

For the two ST450 based stimuli it can be confirmed that the trial with the frontal speaker at 2 m is rated much more reverberant than the trials at 3 and 4.5 m also for a loudspeaker playback.

Auditory distance mapping

The listeners also performed very similar to the binaural test regarding distance mapping.

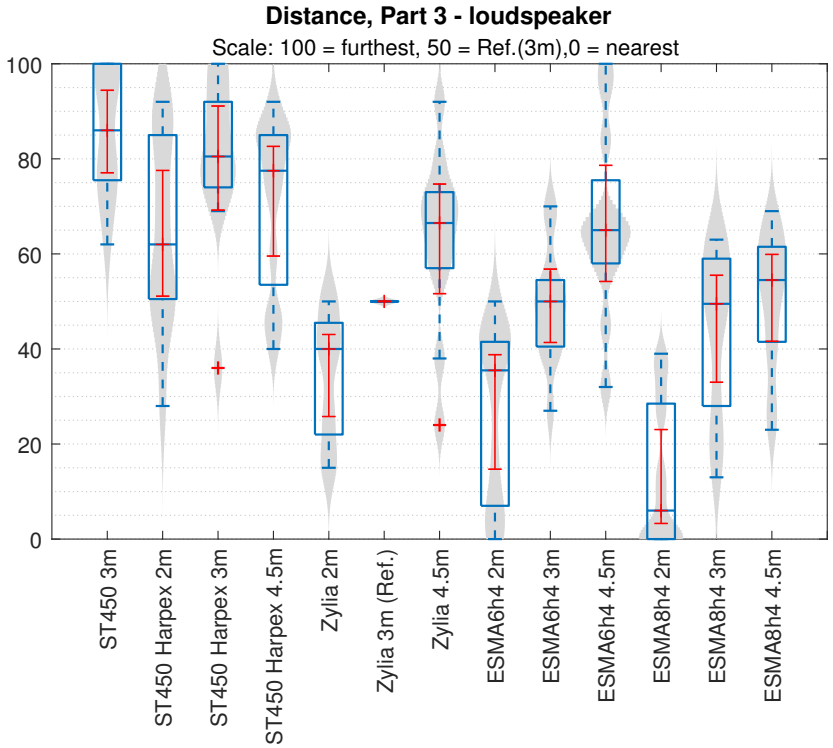


Figure 4.21 – Perceived distances in Part 3 of the binaural test. Blue boxes show the inter-quartile range around the median, blue whiskers extend to the most extreme data points, red crosses show outliers and red whiskers show the 95% confidence interval.

However, in a loudspeaker playback setup distances seem to be imaged better by the ESMA6h4 than the ESMA8h4. The array with six channels complies with the Zylia in contrast to the binaural results. The light gray violin plot shows that within the

loudspeaker-based test distributions tend more towards bi-modality or multi-modality, such as the ST450 Harpex 4.5, Zylia 2m, ESMA6h4 2m and ESMA6h4 4.5m stimulus. It is noticeable that the 2m stimuli are particularly affected, which could be due to an unused scale.

In Figure 4.22 the results of those four participants, which took part in the binaural as well as in the loudspeaker-based experiment, are given. Boxes in blue belong to data of the binaural test, whereas boxes in red show data of the loudspeaker test.

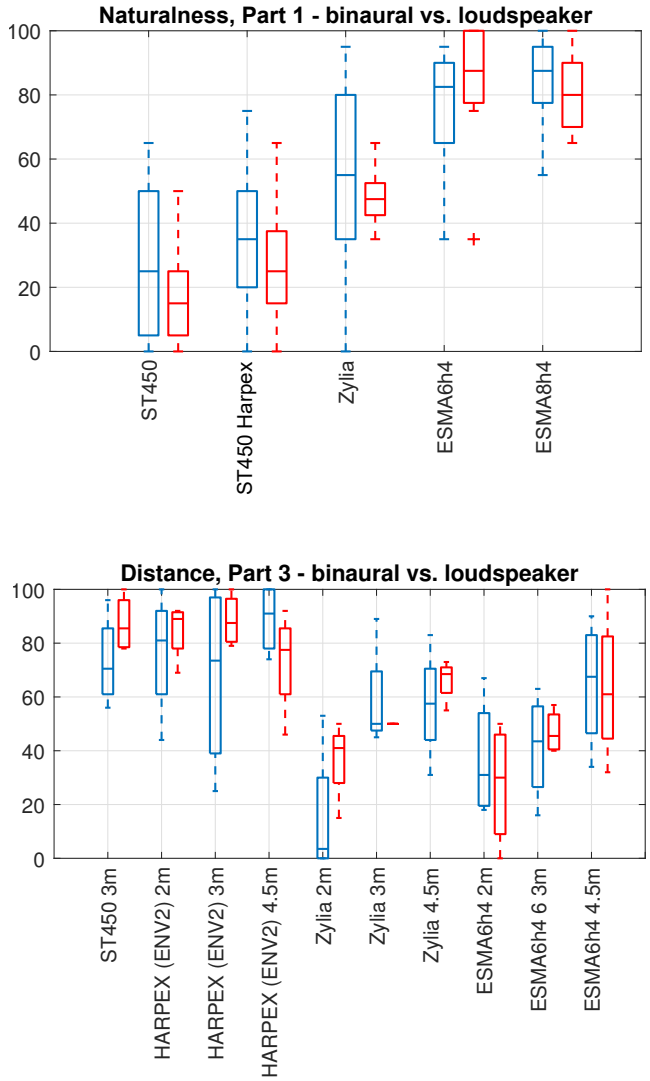


Figure 4.22 – Comparison of results of participants which took part in both experiments. Boxes show the inter-quartile range around the median, whiskers extend to the most extreme data points, crosses show outliers.

The direct comparison shows once more that tendencies of both tests are evident. Additionally the already mentioned difference in spread can be seen. Almost all stimuli tend to be rated rather further away in the loudspeaker based test.

4.2.5 Discussion

The aim of the conducted listening tests was to answer with which 3D-recording technique one comes closest to reality when listening in to the recording in a 91-channel loudspeaker dome. Since no system with a similar high resolution of the target dome was available, the test was conducted binaurally as well as on a small loudspeaker setup to cross-validate the results.

Before summarising the main findings, the difficulties of the 'Plausible Microphone Array Recording' evaluation should be pointed out. In the preparation as well as in the evaluation it was noticed that it is very difficult to design an experiment that compares the plausibility of recordings. Especially if concessions regarding the test vs. target playback setup (91-channel loudspeaker dome) have to be taken in mind.

The main difficulty of the analysis was that the simultaneous query of multiple conditions could not be prevented during the experiment. For example, when evaluating the impression of the room, how should the sound colour of the room be included, but not the array specific frequency responses still remaining after equalization? (It was attempted to adapt those only in direct comparison with the reference noticeable characteristics.) What would have helped would have been to assign priorities to the proposed attributes, so that in case of equality a rating would be easier for the listeners.

Nevertheless, the results allow several interpretations and conclusions: Overall the ESMAs were perceived to be more natural than the Zylia and far more natural than the the two ST450 stimuli. The HARPEX upmix was thereby perceived significantly more naturally on headphones. However, the headphone-based results for the Zylia show no significant difference to the ESMAs if the trial containing the left speaker position is excluded.

A rather remarkable outcome is that the trials for both ESMA-3Ds were rated to be more natural if the sound source is located on-axis with one of the horizontal microphones. The results for the ESMA-3Ds suggest that they reproduce objects more closely on loudspeakers and sound less reverberant but more natural in comparison to the Ambisonics microphones. However, this is completely in line with findings in [Lee19].

The analysis of the spatial impression part is more difficult because the scales were interpreted differently and the results for the individual speaker positions differ significantly. The Zylia came closest to the reference in terms of spatial impression and distance assessment - directly followed by the ESMA-3Ds. A tendency of the ESMA8h4 to sound

drier and closer as the 6-channel version was observed. Whether this is also dependent on the sound source position relative to the microphone position could not be verified.

What is most striking about the ST450 and ST450 Harpex stimuli is that distances and spatial impressions are perceived differently from the others. Thereby, the shortest distance was evaluated as the most reverberant and 4.5 m as close as the 2m position. A possible cause could be that the loudspeaker position at 4.5 m was located relatively close to the rear wall. It is possible that the higher microphone position may have caused the ST450 stimuli to produce a misleading distance image.

Generally the same tendencies and detections could be made for headphones and loudspeakers. Surprisingly the Zylia was rated to be less natural in the loudspeaker test (no reference) compared to the binaural test (including a reference). It is unclear whether the playback mode or the reference is the reason for this. In any case the loudspeaker-based results for Zylia and ESMA-3Ds were more distinct.
Concerning the loudspeaker test the spread is generally lower and the difference between the two ESMA-3Ds and the two ST450 stimuli is negligible. That could be because test participants were allowed to move their head slightly in the loudspeaker based, whereas no head tracking was used for the binaural test.
Other findings from the comparison of the two tests include a clearly better distance mapping for the ESMA6h4 and a less strong effect of the environment parameter (ENV =2) of the HARPEX plugin on speakers. The latter resulted in a more reverberant ST450 Harpex stimulus on headphones.

Overall both ESMA-3Ds were perceived to be the most natural and well-sounding. In contrast to the coincident arrays externalisation worked far better as head-movements did not lead to an unpleasant auditory impression. However, if tonal compromises are accepted, transportability, the USB-field-recorder feature and the comparable low price speak for the Zylia. So especially in remote, sensitive and small recording venues, the Zylia is a nice alternative to the ESMA-3Ds.

4.3 Plausible source embedding

The aim of this second listening experiment is to test how close an automated source embedding with estimated input data from a rough impulse response comes to a reference recording. A MUSHRA test is used to examine the aspects that were also employed to evaluate plausible microphone array recording - naturalness, room impression and distance. The binaural stimuli are created by decoding the Ambisonics signal to virtual loudspeakers and subsequently convolving the resulting with the BRIRs of a KU100 dummy head [4] [CAK18].

Due to the COVID-19 pandemic restrictions that were enacted when writing this thesis, the test was only made available online. Therefore the number of headphones is non-singular but limited to two most common types (Beyerdynamics DT770, DT990 and AKG 271 mk2, 272, 702) regarding frequency response. Since the test was designed by using the DT770, equalization was performed with reference to this model [5]. To consider thinkable effects, the result will also be analysed with regard to the headphones used.

4.3.1 Test design

Within the listening experiment, three different versions of the embedding method are compared with a reference and an anchor stimulus. These six stimuli are described in detail below.

Stimulus 1, Reference
At one lateral position (3m, 90°) and three frontal positions (2m, 3m and 4.5m, 0°) impulse responses were measured via exponential sine sweep in the acoustic target environment with a KU100 dummy head. The BRIRs were convolved with the various sound samples.

Stimulus 2, Embedding 1
For creating this stimulus, as well as the next two stimuli, the sound samples were artificially embedded into the target environment at the same positions as described above. Soleley for this stimulus, unlike described in chapter 3, the first-order early reflections have not been detected separately by a beam. The delays and gains for those first-order wall and ceiling reflections were calculated by using a DOA estimation without any constraint regarding direction or time delay. The resulting early reflections for this stimlus

4. SADIE II database, https://www.york.ac.uk/sadie-project/database.html
5. Database of headphone frequency responses, https://reference-audio-analyzer.pro/en/report/

were then processed with a high-shelving (G= -10dB, Q = 0.5, Fc = 7kHz) and a low-shelving filter (G= -11dB, Q = 0.6, Fc = 90 Hz).By adding Stimulus 2 it is intended to check whether the result is improved by detection of the 1st order reflections by a beam or not.

Stimulus 3, Embedding 2
This stimulus corresponds exactly to the method presented and implemented in chapter 3 without post-processing or frequency-response modification.

Stimulus 4, Embedding 3
Since small differences in the level ratio of direct sound, early and late reflection signal parts had been noticed, the generation of this stimulus aimed at minimizing these difference with regard to the to the reference. To this end, the attempt was to manually compensate for the lack of frequency-dependency or inaccuracies in energy ratios that were due to the acquisition using the flap. Of those three embedding variants, this stimulus is therefore expected to be rated the most similar to the reference.

Stimulus 5, Anchor
An impulse response from the AIR (Aachen Impulse Response) database was used as an anchor [MJV09]. It was recorded with the dummy head by HEAD acoustics in a typically furnished office room via MLS measurement. A volume of 93 ^{3} m and a T_{60} of 0.43 s was specified for the room. For the stimulus position at 4.5 m the measurement data at 3 m was used as from the available data it matched best.

Stimulus 6, Reference 2
Reference 2 corresponds to a dummy-head recording at the given positions. It was played over the same loudspeaker model as the one used for the impulse response measurement of stimulus 1.

The sound samples are on the one hand excerpts from a male voice speech signal as used in the previous evaluation. On the other hand also dry recordings of a telephone ringing sound mixed with typical office table sounds and a copier were re-used in this experiment. Since the implemented method was developed for an embedding in an ambient scene, either a dummy head recording of a simulated office scene or a masking natural noise were used.

The first part consists of only one trial, in which all stimuli should be ranked according to naturalness. The male speaker is located or embedded at 0° at a distance of 3 m and is presented together with a masking natural noise. Without already being influenced

by listening to the stimuli or a given reference, the part is intended to give insight into whether the participants have preferences. The only information provided was that the sound source is located in the front at a distance of 3 m. Another possible statement could be that an embedding variant shows artefacts. Can the subjects tell real from simulated recordings in the absence of a reference?

In the second part the listeners were asked to evaluate the similarity to the reference at different distances and directions. Two different excerpts from the speech sample are presented per position (3m at left, 2m and 4.5m at front). In addition there is a trial with a copier in front at a distance of 4.5m and a trial presenting a ringing telephone in front, as well as typical office noises on the left.

Generally the listeners were notified, that the simulated room as well as the background sound (either an office atmosphere or a masking noise) stays the same for all stimuli and trials. They were also advised that closing the eyes and using the keyboard to toggle between stimuli and reference helps to better concentrate on the virtual room.

Ranking of naturalness, Part 1

Please rank the given stimuli from the most to the least natural embedding - rank 1 (most natural) to rank 6. If you hear no difference at all, it is allowed to give a rank twice and omit the rank below.

You will hear a speaker at the front located at a distance of 3m.

Similarity, Part 2

Please rate the similarity between the reference (dummy head recording) and the presented stimuli.

In case of equivalent overall similarity, the rating should be made according to the following priorities:

1. *Spatial impression*
2. *Distance*
3. *Timbre*
4. *Direction*

4.3.2 Binaural experiment

Post-screening

Since in the first part a subjective rating without repetitions was queried, no test person is excluded.

As no outstanding values in the individual results of part 1 as well as by means of the RMS error of part 2 (figure 4.23), no participant is to be excluded.

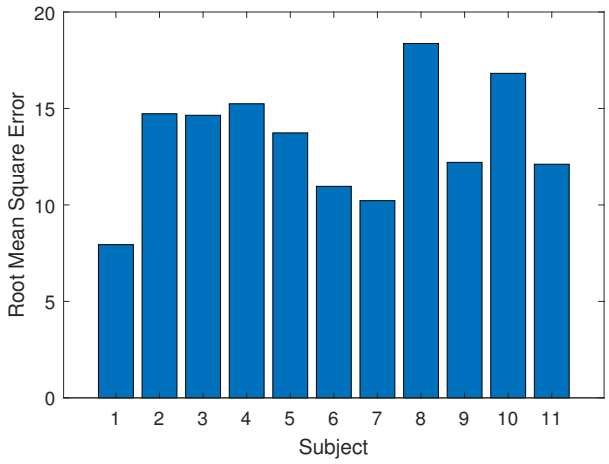


Figure 4.23 – Scaled RMS error for all 11 participants, computed on the evaluations of perceived naturalness. The scaled RMS error is obtained via the mean square error term of the one-way ANOVA (as described in section 4.2.3).

Data

Between April 1 and April 7, 11 sound engineering students and lecturers at the IEM (10 male, 1 female) took part in the online listening test with an average age of 31 years. The experiment took an average of 28 minutes to complete.

The Lilliefors test shows that in part 2 all stimuli are normally distributed except for the anchor. Consequently also pairwise t-test for parametric datasets can be executed.

Results

In the first part, in which the subjective ranking of the naturalness was queried, a discrete rank scale was used. Therefore the significance was examined by a Wilcoxon signed rank test as a pair test, which showed no significant differences for any combination.

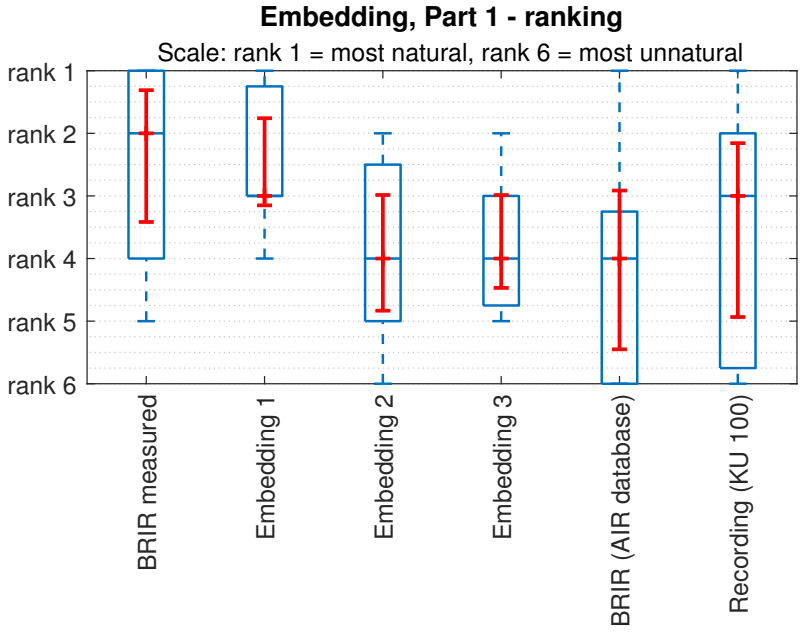


Figure 4.24 – Results of part 1 (*Naturalness ranking*). Whiskers (red) show 95% confidence intervals around medians.

It seems that without a given context or reference, the stimuli cannot be ordered by naturalness. Interestingly, neither the real recording nor the stimulus of the measured BRIRs are clearly evaluated as the most natural stimuli. The dummy head image (see figure 4.25) shows a bimodal distribution indicating that for some listeners the presented snippet was perceived to be more unnatural than all other stimuli.

It is questionable why the BRIR measurement tends to be evaluated more naturally and what caused some of the test subjects to rank the recording with the dummy head as the most unnatural, but it might be due to the sensitivity of the dummy head to the very low frequency end and environmental noises picked up there.

Following the boxplot representation in figure 4.24 it is noticeable that Embedding 1 (generated without correct detection of the 1st order reflections) is evaluated more naturally than the other versions. Embedding 3, which is only a version of Embedding 2 adapted to the reference, is not perceived as more or less natural.

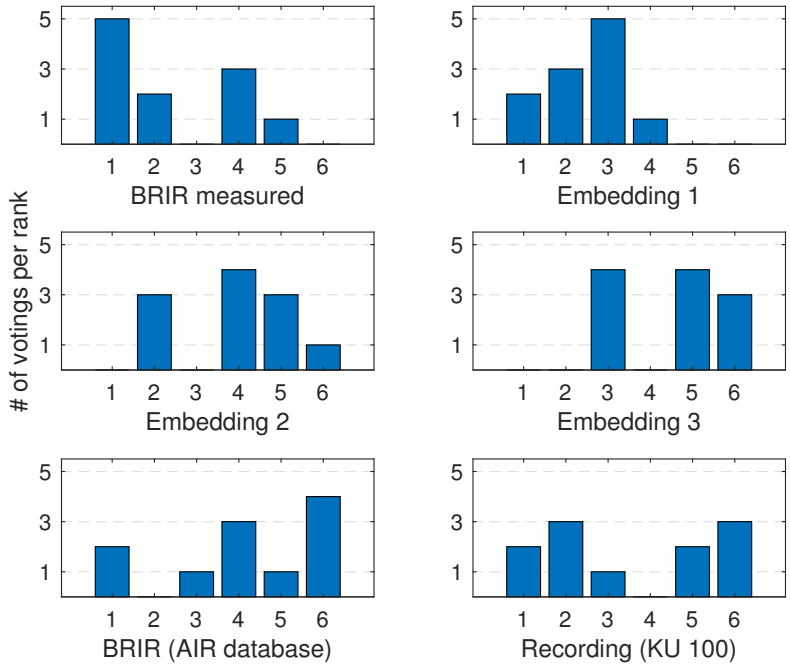


Figure 4.25 – Results of part 1 (*Naturalness ranking*) presented in a barplot diagram.

To generate more information from the data Cliff's δ for ordinally scaled data is also evaluated.

	Emb. 1	Emb. 2	Emb. 3	BRIR (AIR database)	Recording (KU 100)
BRIR measured	0.058	0.562	0.512	0.521	0.388
Embedding 1		0.595	0.587	0.595	0.248
Embedding 2			-0.099	0.132	-0.074
Embedding 3				0.231	-0.041
BRIR (AIR database)					-0.165

Table 4.6 – Table showing Cliff's δ values for Part 1 (*Naturalness ranking*).

Accordingly the effect is negligible when comparing the measured BRIR with Embedding 1, as well as Embedding 2 with 3 and the dummy head recording with Embedding 2 and 3. Contrarily, according to Cliff's δ, the effect is big when comparing the first pair

with Stimuli 2 to 4. Following this analysis, the effect between the BRIR measured with the KU100 and the recording with this dummy head is higher than between Embedding 1 and the recording.

However, in general, the result shows that none of the embedding versions performs significantly less natural than the reference. Listeners appear to find embedding by the presented method credible.

Most often, the BRIR measurement was ranked first for naturalness, confirming that this stimulus is a suitable reference for the second part. How close the embedding comes to reality (represented by BRIR measured) can be seen in the results of part two.

Overall analysis (Similarity)

The Bonferroni-Holm adjusted p-values show that when all trials are polled for an analysis over a greater set of responses, significant differences are found except between embedding stimuli 1 and 2.

In principle, all embedding variants have been evaluated to be similar to the reference in comparison to the anchor. Adjusting energy ratios seems to bring a small but apparently significant improvement.

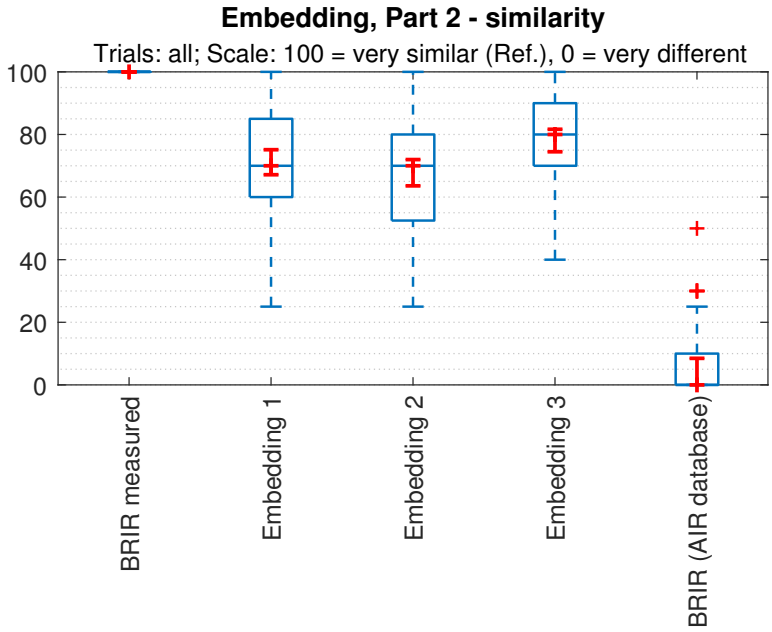


Figure 4.26 – Comparison of all presented trials of Part 2 (*Similarity*). Whiskers show 95% confidence intervals around medians.

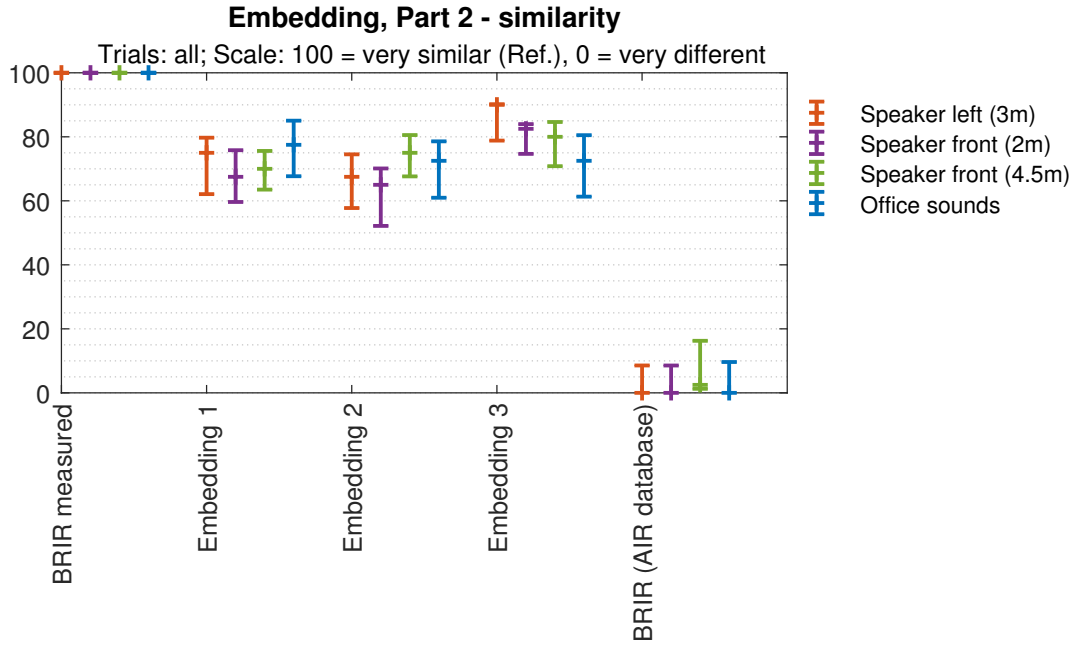


Figure 4.27 – Comparison of all presented trials of Part 2 (*Similarity*). Whiskers show 95% confidence intervals around medians.

Particular scenarios

Figure 4.27 comparatively displays the confidence intervals for each embedding position or sound sample type. For the closer positions of 2 and 3m, the energy tweak of Embedding 3 brings a clear improvement in contrast to the other two trial categories.

4.3.3 Discussion

Overall, the results presented above can be summarised as satisfactory. In the feedback after the listening test, the majority of the subjects gave a good report on the embedding methods developed. The similarity to the reference was evaluated with more than 90 out of 100 possible points for 21% of the answers regarding the embedding stimuli. 12 times 100 points, which is equivalent to mistakably similar, were given.

The answers of the individuals also indicate that if sound samples other than speech had been used, the distinction would have been more difficult. The fact that the trial with the phone and office sounds was relatively often rated as 'very similar' can be attributed to the fact that people are used to hearing subtle differences in voices.

Part 1 allows the interpretation that without reference no difference can be evaluated between a simulation and a recording with the artificial head, which is intended to cor-

respond to human hearing. That Stimulus Embedding 1 seems to sound more natural than the other two Embedding stimuli is surprising. Since the plausibility of the early reflection embedding cannot yet play a strong role without reference, the rough high and low-shelving filtering is held responsible for this.

Embedding Variant 2 without any processing in the frequency domain as well as Variant 3 perform equally similar or more similar. Thus it can be concluded that the accurate detection of the 1st order reflections allows a higher similarity.

However, the result shows that an adjustment of the energy ratios and rough subsequent EQing brings an improvement in the case of Embedding 3. It can therefore be concluded that there might remain further potential in the embedding method. Along with a better decorrelation of the early reflections a standard low-shelving filter like the one used for Embedding 1 could be applied.

In order to assess whether a frequency-dependent parameter estimation and embedding makes a difference or whether the standard EQing is sufficient for a plausible embedding, the listening test would have to be extended to several different rooms.

Beside the frequency dependent parameter estimation there are some more ideas for additional features. As suggested in [AH06] an echo density measure could be estimated to control the room size parameter within the FDN reverb together with the room volume. It would further be interesting to model directional characteristic alignment of sources, as shown in an approach in [WF18].

If sources were dynamically embedded along trajectories in the future, it would need to be worked on room sector interpolation. Moreover, simulation of effects of motion would be of importance, as this is a particular difficulty for hearing-impaired persons [GGH19].

Chapter 5

Conclusion

This thesis investigated ways to take plausible microphone array recordings of everyday life scenes in 3D-audio and to develop an efficient automatic source embedding method to augment these recordings.

The main contribution is a comparative perceptual study on main microphone arrays for 3D-audio recording. Those arrays should be capable of reliably and plausibly reproducing various acoustic environments. On basis of a discussion of several microphone array candidates, a number of them were selected regarding their particular qualities.

Among them were on the one hand coincident arrays with one representative each for first-order and higher-order Ambisonics. Also included was a first-order recording upmixed to higher-order using the HARPEX plugin. On the other hand, two ESMA-3D models with different horizontal resolution were compared in the category of spaced microphone arrays. They were first technically analysed and hence examined using an extended r_E-vector model regarding propagation time.

In the main experiment significant differences between the candidates were found in terms of naturalness, spatial impression and distance. Both ESMA-3Ds were perceived to be the most natural sounding regarding the binaural reference and in contrast to the coincident arrays the externalisation worked better. The fact whether a sound source is on-axis or not with a microphone led to the only rated difference regarding various horizontal resolutions. For the coincident arrays, the Zylia ZM-1 microphone stood out, as it came closest to the reference in terms of spatial impression and distance assessment. The room impression was, however, constantly offset and misleading by the SoundField ST450, nor did the directional sharpening lead to a satisfactory improvement.

It was concluded that among the investigated arrays the ESMA-3D with 6 or 8 horizontal channels is perceptively the best solution for plausible reproduction of an ambient scene. If however transportability and sensitive recording locations become relevant, the Zylia

ZM-1 offers a valuable alternative.

For the augmentation of the recordings by artificial embedding, e.g. of speech, an efficient signal processing model was developed aiming for the highest possible plausibility. The required input parameters were provided by a parameter estimation on the basis of simple flap impulses at the recording venue. In a concluding study on this embedding, it was shown that naturalness can already be achieved in the rendering of the augmenting signals with quite simple means.

As effects of sound motion pose a particular difficulty for hearing impaired people, the extension of the developed model regarding a simulation of moving sounds including source directivity would be of importance regarding future investigations.

Both binaural renderings as well as the original files of the ambient scene recordings are made available via KUG-Phaidra, the online archive of the University of Music and Performing Arts, Graz.

Bibliography

[ABH+10] J. S. Abel, N. J. Bryan, P. Huang, M. A. Kolar, and B. V. Pentcheva, "Estimating room impulse responses from recorded balloon pops," 2010.

[AH06] J. S. Abel and P. Huang, "A simple, robust measure of reverberation echo density," 2006.

[AMD19] A. Ahrens, M. Marschall, and T. Dau, "Measuring and modeling speech intelligibility in real and loudspeaker-based virtual sound environments," *Hearing Research*, vol. 377, pp. 307 – 317, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378595518305598

[BAA+19] F. Brinkmann, L. Aspöck, D. Ackermann, S. Lepa, M. Vorländer, and S. Weinzierl, "A round robin on room acoustical simulation and auralization," *The Journal of the Acoustical Society of America*, vol. 145, pp. 2746–2760, 04 2019.

[Bat16] E. Bates, "Comparing ambisonic microphones: Part 1," *AES*, 2016.

[Ber19] S. Berge, "Acoustically hard 2D arrays for 3D HOA," *AES*, 2019.

[Bla99] J. Blauert, *Spatial hearing - The psychophysics of human sound localization*. The MIT Press, 1999.

[Blo17] M. Blochberger, "FDN reverberation in ambisonics," 2017.

[Blo19] ——, "Signal decorrelation for diffuse sound fields," 2019.

[BSH10] S. K. Bernhard Seeber and E. Hafter, "A system to simulate and reproduce audio visual environments for spatial hearing research," *Hearing Research*, vol. 260, no. 1, pp. 1 – 10, 2010.

[CAK18] D. M. Cal Armstrong, Lewis Thresh and G. Kearney, "A perceptual evaluation of individual and non-individual HRTFs: A case study of the SADIE II database," *Applied Sciences*, 2018.

[Coh88] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. New York: Routledge, 1988. [Online]. Available: https://doi.org/10.4324/9780203771587

[DSP10] F. W. Dirk Schröder and S. Pelzer, "Virtual reality system at RWTH Aachen University," in *Proceedings ICA 2010*. Sydney, Australia: Australian Acoustical Society, NSW Division, 2010, 1 CD-ROM. [Online]. Available: https://publications.rwth-aachen.de/record/118540

[EB+17] M. G. Enda Bates, Sean Dooney *et al.*, "Comparing ambisonic microphones - part 2," *AES*, 2017.

[EFW18] K. Enge, E. Frauscher, and S. Wasserfall, "Richtungsschärfung ambisonischer mikrofonsignale im zeitbereich," Kunstuniversität Graz, Tech. Rep., 2018.

[EKF15] F. P. Eric Kurz and M. Frank, "Comparison of first-order ambisonic microphone arrays," 2015.

[FB12] S. E. Favrot and J. Buchholz, "Reproduction of nearby sound sources using higher-order ambisonics with practical loudspeaker arrays," *Acustica United with Acta Acustica*, vol. 98, no. 1, pp. 48–60, 2012.

[Fra13] M. Frank, "Phantom sources using multiple loudspeakers in the horizontal plane," Ph.D. dissertation, Universität für Musik und darstellende Kunst Graz, 2013.

[FRK05] S. Z. Francis Rumsey and R. Kassier, "Relationships between experienced listener ratings of multichannel audio quality and naive listener preferences," *JASA*, 2005.

[GEH15] G. Grimm, S. D. Ewert, and V. Hohmann, "Evaluation of spatial audio reproduction schemes for application in hearing aid research," *CoRR*, vol. abs/1503.00586, 2015. [Online]. Available: http://arxiv.org/abs/1503.00586

[GGH19] J. L. Giso Grimm and V. Hohmann, "A toolbox for rendering virtual acoustic environments in the context of audiology," *Acta Acustica united with Acustica*, vol. 105, pp. 566–578, 2019.

[GM04] M. M. Gerhard Müller, *Taschenbuch der Technischen Akustik*. Michael Möser, 2004.

[GM17] M. C. Green and D. Murphy, "EigenScape: A database of spatial acoustic scene recordings," *Applied Sciences*, 2017.

[Gri17] S. Grill, "VST-Implementierung Ambisonischer Nachhalleffekte," 2017.

[GS16] M. Giller and B. Stahl, "Directional Audio Coding - Implementierung und experimentelle Bestimmung der optimalen Zeit- und Frequenzauflösung," 2016.

[Int18] *ITU-R BS.2051-2, Advanced sound system for programme production*, International Telecomunication Union Std., 07/2018.

[Jam10] A. Jamalzadeh, "Developing effect sizes for non-normal data in two-sample comparison studies," Durham University, Tech. Rep., 2010.

[JH15] M. S. Jürgen Herre, Fabian-Robert Stöter, "Statistical methods for audio experiments," International Audio Laboratories Erlangen, Tech. Rep., 2015.

[JPJW93] M. M. J. P. Jullien, E. Kahle and O. Warusfel, "Spatializer: A perceptual approach," *AES*, 1993.

[Kur18] E. Kurz, "Efficient prediction of the listening area for plausible reproduction," Master's thesis, Kunstuniversität Graz, 2018.

[Lee13] H. Lee, "Apparent source width and listener envelopment in relation to source-receiver distance," *AES*, 2013.

[Lee19] ——, "Capturing 360° audio using an equal segment microphone array (esma)," in *Audio Engineering Society Convention*, Oct 2019. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=19338

[Let89] T. Letowski, "Sound quality assessment: concepts and criteria," 01 1989.

[LG14] H. Lee and C. Gribben, "Effect of vertical microphone layer spacing for a 3D microphone array," *AES*, 2014.

[Lin14] A. Lindau, "Binaural resynthesis of acoustic environments. technology and perceptual evaluation." Ph.D. dissertation, 06 2014.

[Lin15] ——, "Spatial Audio Quality Inventory (SAQI). test manual. v1.2," Technical University of Berlin. Audio Communication Group, Tech. Rep., 2015.

[LJ19] H. Lee and D. Johnson, "An open-access database of 3D microphone array recordings," *AES*, 2019.

[LJM17] H. Lee, D. Johnson, and M. Mironovs, "An interactive and intelligent tool for microphone array design," in *Audio Engineering Society Convention 143*, Oct 2017. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=19338

[Mar14] M. Marschall, "Capturing and reproducing realistic acoustic scenes for hearing research," Ph.D. dissertation, Technical University of Denmark, 2014.

[MJV09] M. S. Marco Jeub and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," *DSP*, 2009.

[MWK12] J. G. Michael Weitnauer, Michael Meier and G. Krump, "Leistungsrichtige Interpolation von binauralen Signalen," *DAGA*, 2012.

[MZZ20] M. F. Markus Zaunschirm and F. Zotter, "Binaural rendering with measured room responses: First-order Ambisonic microphone vs. dummy head," *Applied Sciences*, 2020.

[ND+19] L. S. Norbert Dillier, Ruksana Giurda *et al.*, "Evaluation of an ILD-based hearing device algorithm using Virtual Sound Environments," *ICA*, 2019.

[OB16] C. Oreinos and J. Buchholz, "Evaluation of loudspeaker-based virtual sound environments for testing directional hearing aids," *Journal of the American Academy of Audiology*, 2016.

[Pel01] R. S. Pellegrini, "Quality assessment of auditory virtual environments," 2001.

[PM+13] S. A. Pauli Minnaar *et al.*, "Reproducing real-life listening situations in the laboratory for testing hearing aids," *AES*, pp. 261–270, 01 2013.

[RKCS06] J. Romano, J. Kromrey, J. Coraggio, and J. Skowronek, "Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and other surveys?" in *Florida Association of Institutional Research*, 2006, pp. 1–3.

[RL17] H. Riaz and H. Lee, "Multichannel microphone array recording for popular music production in virtual reality virtual realityï¿½," in *Audio Engineering Society Convention*, Oct 2017. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=19338

[Roc97] D. Rocchesso, "Maximally diffusive yet efficient feedback delay networks for artificial reverberation," *Signal Processing Letters, IEEE*, vol. 4, pp. 252 – 255, 10 1997.

[Rud19] D. Rudrich, "Iem plugin suite," 2019. [Online]. Available: https://plugins.iem.at/

[Rum02] F. Rumsey, "Spatial quality evaluation for reproduced sound:terminology, meaning, and a scene-based paradigm," *Audio Eng. Soc., Vol. 50*, 2002.

[SB13] N. Schinkel-Bielefeld, "Audio quality evaluation by experienced and inexperienced listeners," *JASA*, 2013.

[SH17] S. J. Schlecht and E. A. Habets, "Accurate reverberation time control in feedback delay networks," 2017.

[SH19] S. Schlecht and E. Habets, "Dense reverberation with delay feedback matrices," 10 2019.

[SKWD19] L. S. R. Simon, A. Kegel, H. Wüthrich, and N. Dillier, "3D localization of speech by mildly and moderately hearing-impaired persons in ecological environments," in *Proceedings of the 23rd International Congress on Acoustics*. EAA - ICA, September 2019. [Online]. Available: https://doi.org/10.5167/uzh-182063

[SM09] L. Simon and R. Mason, "Localization curves for a regularly-spaced octagon loudspeaker array," *AES Journal*, 2009.

[SMB06] J. D. Sebastien Moreau and S. Bertet, "3d sound field recording with higher order ambisonics – objective measurements and validation of a 4th order spherical microphone," vol. 1, 01 2006.

[ST12] P. Seetharaman and S. Tarzia, "The hand clap as an impulse source for measuring room acoustics," *AES*, 2012.

[Too85] F. E. Toole, "Subjective measurements of loudspeaker sound quality and listener performance," *AES*, 1985.

[TWE14] S. v. d. P. Torben Wendt and S. Ewert, "Efficient synthesis of perceptually plausible binaural room impulse responses," *DAGA*, 2014.

[Vor07] M. Vorländer, *Auralization: Fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Springer Science & Business Media, 2007.

[WD99] M. Williams and G. L. Dû, "Multichannel sound recording - Multichannel Microphone Array Design (MMAD)," *Audio Engineering Society*, 1999.

[Weg20] K. Wegler, "Modellierung der Echounterdrückung unterschiedlicher Reflexionseigenschaften," Master's thesis, Universität für Musik und darstellende Kunst Graz, 2020.

[WF18] F. Wendt and M. Frank, "On the localization of auditory objects created by directional sound sources in a virtual room," *Tonmeistertagung*, 2018.

[Wil10] M. Williams, "Multichannel microphone array design (mmad)," *Audio Engineering Society*, 2010.

[Wil13] ——, "The psychoacoustic testing of the 3-D multiformat microphone array design and the basic isosceles triangle structure of the array and the loudspeaker reproduction configuration," *AES*, 2013.

[WT02] H. Wittek and G. Theile, "The recording angle - based on localization curves," *AES*, 2002.

[ZF12] F. Zotter and M. Frank, "All-Round Ambisonic Panning and Decoding," *AES*, 2012.

[ZF19a] ——, *Ambisonics - A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer, 2019.

[ZF19b] ——, "Reproducibility of stereophonic amplitude-panning curves," *DAGA*, 2019.

Appendices

Appendix A

Recorded scenes

Office

The acoustic scenery is composed out of

- a continuous and steady computer fan noise
- keyboard tapping and and mouse clicking sounds
- indistinct chatter
- footsteps on wooden floor
- doorbell, photocopier, coffee machine
- damped sound of cars passing by

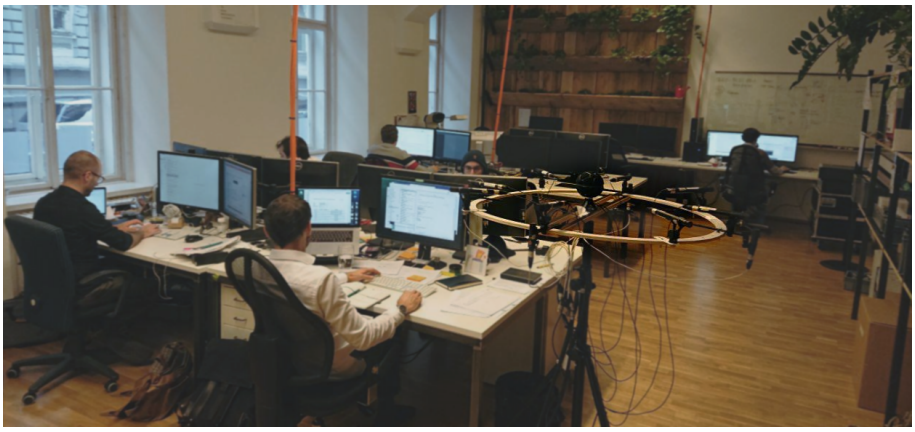
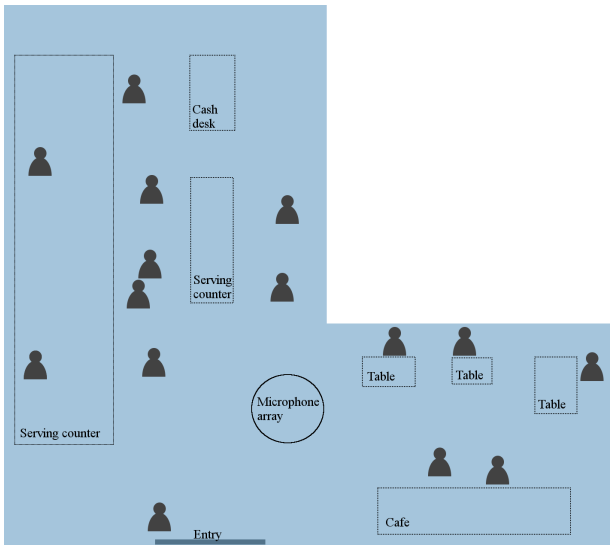


Figure A.1 – The open-plan office contains 15 workspaces, during the recording session half of the employees were present. Sound atmosphere: quiet, low complexity of sound sources, low reverberation time (the ceiling was treated with acoustic absorbers). Room dimensions are l=15m, w=6m, h=4m.

Restaurant

— people chatting
— clattering of crockery
— coffee machine
— roller carriage passing by
— noise floor of refrigerator, deduction, stove,..



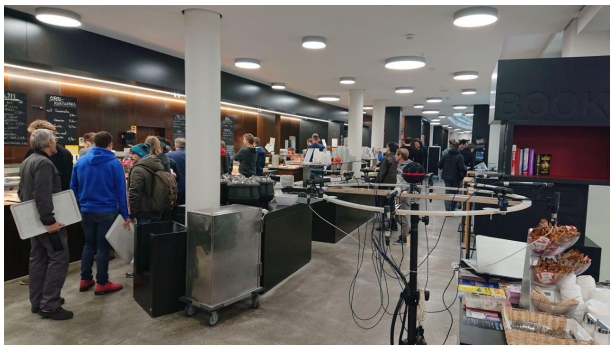


Figure A.2 – Canteen recording venue. Sound atmosphere: busy, high complexity of sound sources, relatively high amount of people, high background noise level. The canteen at the campus of the Technical University Graz was chosen because it was possible to place the array in a way no table conversations are recorded. It was positioned on the knee of the L-shaped room.

Shopping mall

— a lot of babble noise
— people passing by
— coffee shop
— children running and shouting
— cash desk

(a) Shopping mall venue.

(b) Shopping mall array.

Figure A.3 – Sound atmosphere: busy, high reverberation time and big room dimensions, high amount of people, high complexity of sound sources. A permission to record was approved by the shopping mall 'Citypark' in Graz.

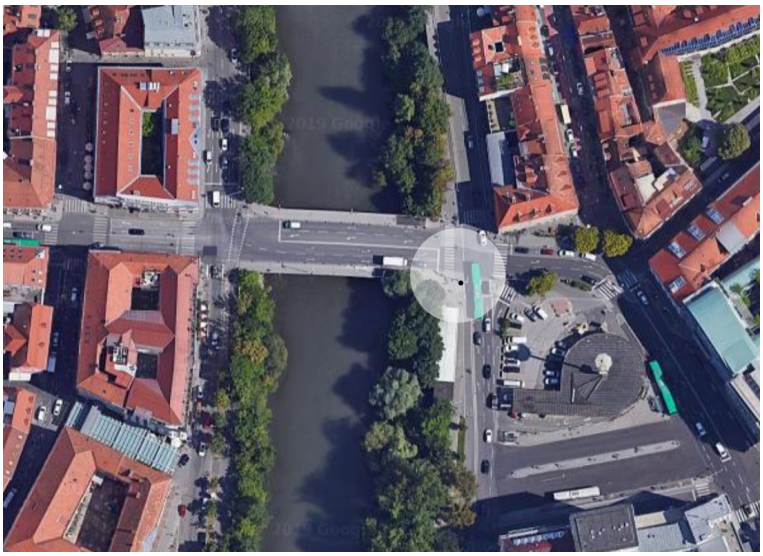
Playground

— children running, chatting, shouting
— distant traffic
— birds
— table tennis
— typical playground sounds (slide, sandbox,..)
— footsteps on leaves (due to the recording date in early november)

Figure A.4 – The playground is located in a park in the center of Graz. The recording venue is marked in the map.

Crossing

— cars and busses passing and stopping at the red light
— bicycles passing and stopping
— pedestrians
— river flowing slowly (Mur)



(a) Crossing venue marked on a map.

(b) Crossing venue.

Figure A.5 – The recording took place at 6pm at the 'Andreas-Hofer-Platz' in Graz.

Trainstation

- trains arriving and leaving
- trains passing
- people wating and chatting
- announcements

Figure A.6 – Trainstation venue at the main railway station Graz.

Trainstation hall

- people passing
- babble
- trolleys, ticket vending machines
- announcements

Figure A.7 – The scene was recorded in the hall of the main railway station of Graz at 5 pm. Sound atmosphere: high reverberation level.

Public space

- busses and trams arriving and leaving
- people passing, waiting
- distant traffic
- tram bell, drinking fountain

Figure A.8 – The recording took place at 7 pm at Jakominiplatz, which is the main public transport hub in Graz.

Park

- strollers, cyclists
- birds
- distant traffic
- children

Figure A.9 – The recording venue within the park is situated in the heart of Graz inbetween the historical center and a main road.

Appendix B

Stimuli recording protocol

Recording date: 27.2.- 2.3.2019.

Location: Haydngasse 10,8010 Graz.

Recording setup: 2 virtual interfaces (as possible on a Mac) were necessary in order to record via MADI and Zylia USB interface simultaneously.

#	Item
1	RME Madiface; MADI Interface
1	Andiamo.MC; Preamp
7	AKG C480; Cardioid
1	AKG C451; Cardioid
4	Neumann KM150; Supercardioid
1	Neumann KU100; dummy head
1	SoundField ST450 MKII (incl. big stand)
1	Zylia (incl. short threaded bolt)
1	NTI measurement microphone
9	Genelec 8020 loudspeaker (3 stands)
1	Wooden ring (3 stands)
8	Short stereo bars
1	distance/angle laser tracker
1	Starting flap

Table B.1 – Equipment list

Recorded configurations and sources:

	6,(7) and 8-channel ESMA-3D
d = 1.24 m	Simulated office scene (SOS) SOS + speech sample @ loudspeaker positions Sweeps @ loudspeaker pos. Sweeps @ SOS loudspeakers pos. Flap impulse @ room sectors
d = 1.38	Simulated office scene (SOS) SOS + speech sample @ loudspeaker pos. Flap impulse @ room sectors
d = 0.8	Simulated office scene (SOS) SOS + speech sample @ loudspeaker pos. Flap impulse @ room sectors Speech sample @ loudspeaker pos. Claps @ room sectors

This led to a total of 73 38-channel one-minute audio files. (Originally also a 7-channel ESMA array was included.)

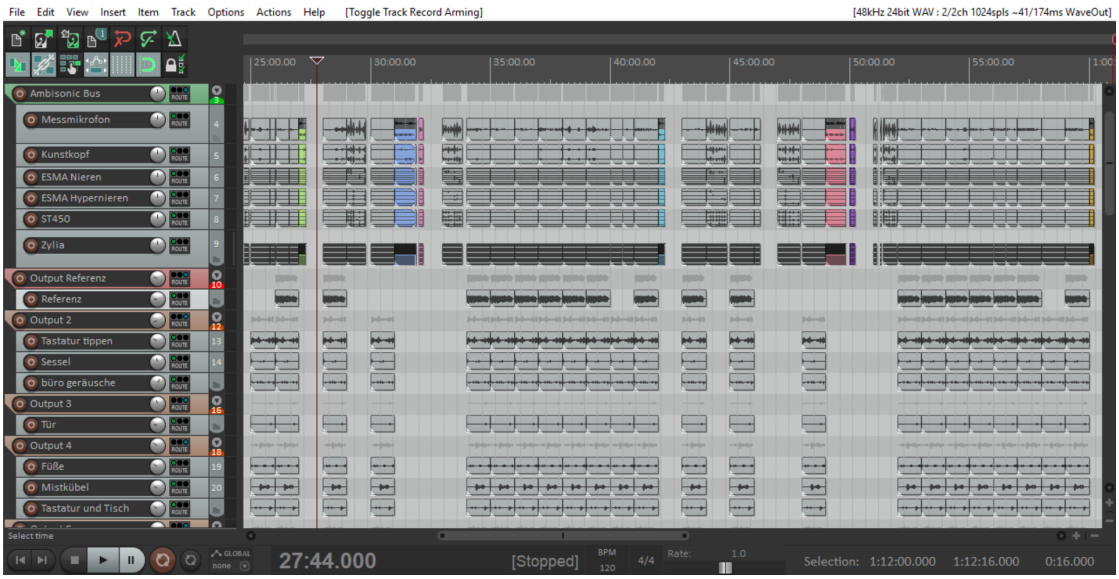


Figure B.1 – Reaper stimuli recording session

Appendix C

Statistics

#	Stimulus	h_{part1}	h_{part2}
1	ST450 Harpex	0	0
2	ST450	0	0
3	Zylia	0	0
4	ESMAh 6	0	0
5	ESMAh 8	0	0
6	KU100	-	1

Table C.1 – Results of normal distrubtion tests with Lilliefors and Jarque-Bera-test for part 1 and part 2 (binaural). '0' indicates that the H_0 (data comes from the normal family) is applicable, whereas '1' indicates that H_0 is rejected ($\alpha = 0.05$).

#	Stimulus	h_{part1}	h_{part2}
1	ST450	0	1
2	ST450 Harpex	1	1
3	Zylia	1	0
4	ESMA6h4	1	0
5	ESMA8h4	0	0

Table C.2 – Results of normal distrubtion tests with Lilliefors and Jarque-Bera-test for part 1 and part 2 (loudspeaker)

#	Stimulus	h_{part3}
1	KU100 2m	0
3	KU100 4.5m	0
4	ST450 Harpex 2m	0
5	ST450 Harpex 3m	1
6	ST450 Harpex 4.5m	0
7	ST450 2m	1
8	ST450 3m	0
9	ST450 4.5m	0
10	Zylia 2m	0
11	Zylia 3m	1
12	Zylia 4.5m	0
13	ESMAh 6 2m	0
14	ESMAh 6 3m	0
15	ESMAh 6 4.5m	0

Table C.3 – Results of normal distribution tests with Lilliefors and Jarque-Bera-test or part 3 (binaural)

Abbreviations tables C.4-5:

K	... KU100 dummy head
S	... ST450
SH	... ST450 Harpex
Z	... Zylia
E6	... ESMA6h4
E8	... ESMA8h4
(R)	... reference
number	... distance in m

	K4	S2	S3	S4	SH2	SH3	SH4	Z2	Z3	Z4	E6 2	E6 3	E6 4
K2	0	1	1	0	1	1	1	1	0	0	0	0	0
K4		0	0	0	0	0	0	1	0	0	0	0	0
S2			0	0	0	0	0	1	0	0	0	0	0
S3				0	0	0	0	1	1	1	1	1	1
S4					0	0	0	1	0	0	0	0	0
SH2						0	0	1	0	0	1	1	0
SH3							0	1	1	0	1	1	0
SH4								1	0	0	1	0	1
Z2									1	1	0	1	0
Z3										0	0	0	0
Z4											0	0	0
E6 2												0	0
E6 3													0

Table C.4 – Table showing the rejection of the hypothesis H_0 (significance) for Part 3 (binaural). '0' indicates that H_0 is applicable and there is no significant difference, whereas '1' indicates that H_0 is rejected by means of $\alpha = 0.05$.

	SH2	SH3	SH4	Z2	Z3(R)	Z4	E6 2	E6 3	E6 4	E8 2	E8 3	E8 4
S3	0	0	0	1	1	0	1	1	0	1	1	0
SH2		0	0	0	0	0	0	0	0	1	0	0
SH3			0	1	0	0	1	0	0	1	0	0
SH4				1	0	0	1	0	0	1	0	0
Z2					0	0	0	0	1	0	0	0
Z3(R)						0	0	0	0	1	0	0
Z4							1	0	0	1	0	0
E6 2								0	0	0	1	0
E6 3									0	1	0	0
E6 4										1	0	0
E8 2											1	0
E8 3												0

Table C.5 – Table showing the rejection of the hypothesis H_0 (significance) for Part 3 (loudspeaker). '0' indicates that H_0 is applicable and there is no significant difference, whereas '1' indicates that H_0 is rejected by means of $\alpha = 0.05$.