EUROPÄISCHE UNION DER HÖRAKUSTIKER e. V.

Förderpreis 2020

Evaluation of spatial sound quality in ecologically valid scenarios

(Evaluation räumlicher Aspekte von Klangqualität in ökologisch validen Szenarien)

Masterarbeit

Verfasser: Stephan Müller
Erstprüfer: Prof. Dr. Tim Jürgens
Zweitprüfer: Prof. Dr. Jürgen Tchorz & Dr. Volker Kühnel
Datum der Abgabe: 23. September 2019

EUHA
Europäische Union der Hörakustiker e.V.

Masterarbeit im Studiengang Hörakustik und audiologische Technik
Institut für Akustik
Universität zu Lübeck

Herausgeber: Europäische Union der Hörakustiker e. V.
Neubrunnenstraße 3, 55116 Mainz, Deutschland
Tel. +49 (0)6131 28 30-0
Fax +49 (0)6131 28 30-30
E-Mail: info@euha.org
Internet: www.euha.org

Alle hier vorhandenen Daten, Texte und Grafiken sind urheberrechtlich geschützt. Eine Verwertung über den eigenen privaten Bereich hinaus ist grundsätzlich genehmigungspflichtig.

© EUHA 2020

Abstract

Many hearing aid features influence the auditory perception of space. Methods to assess the quality of spatial perception in ecologically valid scenarios are highly desired. Using Higher Order Ambisonics reproduction, five scenes in two rooms are recorded that each differ in their ecological validity. Four tasks are developed to assess the subjective spatial quality in distinct dimensions. The tasks consisted of subjective verbal reports, comparative ratings of spatial attributes and the recreation of the presented scenes with figurines on a grid. Eleven expert listeners from Sonova participated in this experiment. The results indicate suitability and reliability of the methods to assess spatial perception in distinct dimensions. Additionally to quality measures of the created tasks and scenes, the results contain valuable information about spatial perception in virtual acoustic environments in general. Further research is needed to facilitate the tasks for the applicability for older hearing impaired subjects.

Zusammenfassung

Viele Funktionen in modernen Hörgeräten beeinflussen die akustische Wahrnehmung des Raumes. Methoden zur Beurteilung der Qualität der räumlichen Wahrnehmung in ökologisch validen Szenarien sind sehr gefragt. Mittels Higher Order Ambisonics werden in zwei Räumen fünf Szenen aufgenommen, welche sich in ihrer ökologischen Validität voneinander unterscheiden. Vier Aufgaben werden gestellt, um die wahrgenommene räumliche Qualität in verschiedenen Teilaspekten zu beurteilen. Die Aufgaben bestehen aus verbalen Berichten der Wahrnehmung, vergleichender Bewertung vorgegebener räumlicher Attribute und der Nachbildung der Szene mit Figuren auf einem Raster. Elf Probanden aus einer Expertengruppe der Sonova nahmen an dem Experiment teil. Anhand der Ergebnisse werden die Methoden als zuverlässig und geeignet für die Beurteilung räumlicher Wahrnehmung in verschiedenen Teilaspekten bewertet. Zusätzlich zu Qualitätsmaßen über die gestellten Aufgaben und Szenen, enthalten die Ergebnisse wertvolle Erkenntnisse über räumliche Wahrnehmung von virtuellen akustischen Szenarien im Allgemeinen. In weiteren Untersuchungen müssen die Aufgaben erleichtert werden für eine bessere Eignung mit älteren Hörgeräteträgern zu erzielen.

Acronyms

HOA Higher Order Ambisonics

WFS Wave Field Synthesis

VBAP Vector Based Amplitude Panning

ITD Interaural Time Difference

ILD Interaural Level Difference

RIR Room Impulse Response

SNR Signal-to-Noise-Ratio

SPL Sound Pressure Level

UDP User Datagram Protocol

GUI Graphical User Interface

DRR Direct-to-Reverb-Ratio

PCA Principal Component Analysis

PC Principle Component

CDF Cumulative Distribution Function

SCN Scenes (factor)

ROM Rooms (factor)

SUB Subjects (factor)

REP Repetitions (factor)

SPK Speakers (factor)

List of Tables

List of Figures

Contents

1 Introduction

Auditory perception of space plays an important role in our everyday lives. While most physical parameters of a sound emitting object are effortlessly captured visually, the auditory spatial information is influenced by multiple characteristics of itself and its surrounding space. Besides the function of triggering the attention on potential dangers or desired information of objects that are outside of the visual field, spatial hearing allows to distinguish unwanted from wanted signals in complex listening scenarios. The skill of identifying sound sources and analysing their relevance for the listener is performed unconsciously by the auditory system at all times. Most information about the location of sound sources is provided by interaural time and level differences and the spectral shape of the signals by the auricle [1].
When the hearing capability deteriorates over time, facilitating effects by auditory spatial perception in complex listening scenarios become more relevant. Increasing difficulties in such situations result in the affected persons to isolate themselves from their social environment. According to a publication from the U.S. National Center for Health Statistics, approximately 15 % of American adults are affected by hearing loss [2]. The application of hearing aids focuses on restoring speech intelligibility in demanding situations, while concomitant influences on spatial perception are often neglected. Such influences have been demonstrated to be induced by hearing aids and some of their features. Adaptive directional microphones for example have shown to impair horizontal localization accuracy [3]. As another example, temporal processing and intensity of compression schemes in hearing aids have been observed to cause more diffuse images of sound sources, create the perception of motion or split images of sound sources [4][5]. Examples like this demonstrate the need for measures to assess the spatial perception with hearing aids.

In sound reproduction industries, it is accepted that physical properties of sound reproducing systems do not directly indicate their perceptive properties. In order to assess the perceptive nature of a reproduction system, it has to be tested and evaluated by human listeners. The matter becomes more important, when sound quality shall be determined. Sound quality itself is a generic term that includes numerous properties that aim to characterize the nature of sound events, such as shrillness or loudness. Corresponding physical measures can be captured using microphones and audio analysers, though subjective perception of such attributes can only be measured with human listeners. Furthermore, the subjective or psychoacoustic properties of the attributes do not imply a certain preference in the listeners. The aim of the assessment of sound quality is to find out the relationship of the technical, subjective and preference measures in order to optimize the products accordingly [6]. Sound quality assessment with hearing aids has experienced increasing attention over the past years, as improved sound quality has shown to reduce prevalence of unworn hearing aids and to improve the perception of hearing aid benefits [7].

An important subcategory to sound quality is the broad range of spatial audio characteristics. In room acoustics, spatial aspects of sound quality have been assessed for some time, for example in the conception of concert halls. In this field, the influence of the materials of the ceilings and walls or the positions of loudspeakers on music instrument perception are assessed. A modern field of application for spatial sound quality assessment lies in virtual audio reproductions, e.g. for applications of gaming in virtual reality. Here, the aim of the audio reproduction technique is to create immersive and plausible spatial sensations. In the field of hearing aid research, auditory perception of space has mostly been evaluated in isolated disciplines such as localization accuracy [3] or distance perception [8]. Outside of laboratory settings, the perception of position and distance of sound sources is influenced by the characteristics of its surrounding space and their positions relatively to adjacent obstacles. Additionally to the spectral or interaural cues, much information about the objects is derived from sensory-motor feedback. Ecological validity of experiments that aim to assess spatial quality of hearing aids means that such mechanisms should be included.

Several techniques exist to reproduce realistic acoustic scenarios in laboratory settings. One such technique is Higher Order Ambisonics (HOA) reproduction. The advantage regarding the presentation of spatial aspects of a scenario is that the specific room characteristics and the relationship between sender and receiver are preserved. Furthermore, the technique allows listeners to make use of sensory-motor feedback as spatial cues by turning their head within certain limits. The technique is also established and has been tested regarding applicability with hearing aids [9].

In this work, a method is developed and evaluated in order to measure spatial sound quality in ecologically valid scenarios. For the presentation of sound material that contains spatial information, different scenes are created using HOA. Four tasks are conducted in order to assess the spatial sound quality of the test material. The tasks include verbal and non-verbal elicitation techniques. The demand on the experiment is high ecological validity and good applicability for hearing aid feature assessment.

2 Background

2.1 Spatial Hearing

Spatial hearing is a skill that comprises several abilities. This section shall give an overview of the different aspects of spatial hearing and the mechanisms that enable to perceive these aspects.

An essential skill of spatial hearing is the localization of sound emitting objects in the environment. The source direction, i.e. the angle of an auditory object around the listener is determined by three acoustic cues: The Interaural Time Difference (ITD), the Interaural Level Difference (ILD) and direction-dependant spectral shaping by the auricle [1].
The energy of a sound wave that spreads in space disperses with increasing distance to the source. When hitting the head of a listener, the sound waves are partly absorbed and partly reflected. This results in an acoustic shadow, the size of which depends on the ratio of wavelength to head diameter. Both effects result in differing sound pressure levels on both ears for sound sources coming from the side, i.e. the ILD. The sensitivity for ILD cues is highest above $2\,\mathrm{kHz}$.
As sound waves move at constant velocity through space, the distance between two ears results in a temporal delay for the contralateral ear for sounds coming from the side, i.e. the ITD. In the auditory system, the basic mechanism to extract temporal information from the sound signals is called phase locking. Specialized neurons synchronize their firing rate to the phase of the signal. The signals from the neurons of both ears come in together in so called coincidence detector neurons. The activity of these neurons is only exhibited when the firing rate of both signal paths are aligned. The source direction is then decoded by which population of coincidence detector neurons is firing. The firing rate of phase locking is limited up to frequencies of $1.3\,\mathrm{kHz}$ [10][11].
Depending on the spectral shape of the signal, listeners give higher weight to the cue that provides more information. For low-pass filtered signals, the ITD cues are given more weight, for high-pass filtered signals, the ILD cues become dominant. For broadband stimuli, the ITD cues are given more weight. The relationship of both interaural cues is described in the duplex theory [12].
Both interaural cues form the necessary information in order to locate sounds in the horizontal plane. When the task is to distinguish sound sources from the front and back, the interaural cues become ambiguous. Without further information, both cues form a cone of confusion, meaning that the cues for one sound source are not clearly encoding the position along a cone around the frontal axis of the head. This includes the absence of information about the elevation of the sound source. As sound waves are being partly reflected and absorbed at the upper body and especially the auricle, the spectrum of the signals is shaped by notches and peaks. The shape of this spectral pattern is dependant on the source direction of the object. Being able to decode the spectral cues resolves the issue of the cone of confusion. Because the shape of the auricle is

individual and changes through age, the auditory system is capable of adapting to changes in the anatomy [13].
In addition to the directional localization of sound emitting objects, human listeners are able to estimate the distance of sound sources. The distance perception comprises the detection of multiple characteristics in the signal. A primary acoustic property of a sound source that determines its distance is the intensity level. With increasing distance, the intensity level of an object decreases. With point-like sources in free-field or anechoic environments, the decrement in intensity is 6 dB per doubling in distance. For lateralised sounds, the level differences between the ears is strongly dependant on their distance. Closer sounds produce larger ILDs, distant sounds produce smaller ILDs. In non-anechoic environments, parts of the sound signal reaches the receptor directly, and other parts are reflected and absorbed by the surrounding surfaces. The ratio of the energy of the direct and indirect sound path contains information about its distance in rooms and is referred to as Direct-to-Reverberant Energy Ratio (DRR). When the sound source is farther away, the absorbent properties of the transmitting medium shapes the spectrum of the signal [14].

The perception of space is a multimodal ability, meaning more than just the auditory input contributes to the characterization of the environment. Outside of laboratory settings, the scenarios in which we find ourselves are often complex and provide a flood of information that contributes to the spatial impression. In such situations, the motion of sound sources provides additional information about their location in space. Furthermore, slight head movements of the listener shows to facilitate localization. While the head moves, the muscular system of the neck and the vestibular system supplies the listener with additional information about the own orientation in space [15][16].
Audio-visual integration plays an important role in the perception of the environment. The location of auditory objects in our surroundings is greatly biased by the visual information. The so called *Ventriloquist Effect* describes a bias in localization towards visual objects when both visual and auditory information are incongruent. This effect can last for several minutes after removal of the visual input [17][18]. It has also been demonstrated that the visual width of a sound emitting object influences the perceived auditory source width of that object. Furthermore, the visual impression of a room induces expectations towards perceived room acoustical parameters [19]. It is also observed that the auditory distance estimation is influenced by visual stimulation [20]. Despite the high influence of surrounding surfaces on the perception of a sound source, few studies investigated how listeners characterise the properties of the surrounding space itself. It could be demonstrated, that the perception of room size is strongly dependant on the reverberation time and the source-receiver distance [21][22].

In the context of this work, the ability to detect changes in the facing angles of sound emitting

objects shall be addressed. In our everyday lives, the facing angle is a crucial cue to distinguish speakers that address the listener from those that do not. It has been demonstrated that the estimation of the source facing angle is more accurate when it sounds from in front of the listener than from the sides. The estimation also becomes less accurate the further the source is facing towards the sides [20]. When a speaker is facing away from the listener, the signals spectrum is shaped by a low-pass filter, which humans are able to detect [23]. The relevance in our everyday lives is demonstrated as the facing angle of concurrently active speakers influences the speech perception of target speakers [24].

2.2 Virtual Acoustic Environments

It has been pointed out that spatial hearing in everyday lives underlies influences from multiple sensory modalities and acoustic parameters along the signal paths. Consequently, it would be ideal to evaluate the spatial hearing ability in situations that include such influences. However, in order to draw general conclusions about isolated aspects of spatial hearing, it is required to create experiments in controlled and reproducible environments. In such laboratory settings, the influences or biases of information that are not related to auditory cognition are minimized, while it is still important not to dissociate too much from ecologically valid scenarios. The development of hearing aids has brought forth features that adapt to changes in the environments, e.g. that change the directivity pattern in demanding situations. In order to evaluate the benefits from such features, it is important to recreate demanding situations in laboratory settings that are perceived as plausible by the listener while also presenting realistic signals for the hearing aid.

Many technologies that aim to reproduce realistic acoustic scenarios in laboratory settings have been developed. The term virtual acoustics refers to the creation of sound events that are perceived from physically non-existent sources. Alternatively to virtual acoustics, laboratory scenarios can be created using the discrete loudspeakers as existent sound sources. This approach is limited, as the number of loudspeakers determines the potential source positions of both target signals and ambient sound sources and the room acoustic influences on perception are bound to the measurement room. Recreation of more complex scenarios that include changes to e.g. room acoustic parameters are not implemented at reasonable expense. The creation of virtual acoustic environments principally resolves such limits to the complexity of acoustic scenarios. Several methods to create such virtual environments exist, that each come with individual advantages and disadvantages. Virtual acoustic environments can be created by recording real-life scenarios or by developing artificial scenarios. The choice of the method has to be evaluated based on the intention of the experiments. In this experiment, the demand on the scene presentation is to be suitable for hearing aid evaluation [25].

One such technique that preserves the individual relationship of sound source and listener (including reflections of pinna, head and torso) is to compute head-related transfer functions (HRTF). This transfer function can be convolved with any desired signal. The disadvantage of this technique is that the material has to be presented over headphones, which complicates the usage of sensory-motor or vestibular feedback as spatial cues. Also, hearing aid algorithms can only be tested as a part of the processed signal, which precludes the influences of individual acoustic ear coupling or usage of own hearing aids [26].
Approaches that make use of multiple loudspeakers are Vector Based Amplitude Panning (VBAP), Higher Order Ambisonics (HOA)[27] and Wave Field Synthesis (WFS) [28]. VBAP is a method, where virtual sound sources are created in between existing loudspeakers by geometric weighting of the signal gains for the contributing speakers. Complex acoustic scenes could be created by applying several such virtual sound sources, though the method does not take the influence of reverberation on many aspects of spatial hearing into account. This precludes the creation of scenes in different rooms apart from the measurement chamber [27]. HOA is a technique that aims to physically recreate the sound field within the limits of a so called «sweet spot», while WFS aims to recreate the whole sound field. The plausibility of the sound fields of both methods is limited by the number of available loudspeakers. Though in the case of WFS, the sound field's accuracy is only provided if the distance between the loudspeakers is less than half the wavelength of the maximal occurring frequency [27]. The amount of necessary loudspeakers is thus considered impractical to this experiment. An advantage of the HOA approach over WFS is that the technique of both recording and playback of virtual acoustic scenarios is established and well explored.

HOA describes a mathematical method for the decomposition and resynthesis of a sound field. In practice, a real sound field that is recorded with an array of microphones is encoded (decomposed) into spherical harmonics that represent the directional information of the sound field. The encoded signal can then be decoded (resynthesized) for a specific array of loudspeakers. Once acquired, the encoded HOA signals are not specific for any loudspeaker array, thus can be transferred between different laboratory settings. The order of the Ambisonics is determined by the number of decomposed spherical harmonics. The first order signal consists of four channels, while the first represents the omnidirectional information and the others contain the dipole-shaped information along the X-, Y- and Z-direction. With each further Ambisonics order, a set of spherical harmonic channels add to the previous ones. Second order Ambisonics consists of 9 channels, third order of 16 channels, fourth order of 25 channels and so on. The order of Ambisonics determines the size of the sweet spot. While the size of the sweet spot grows by the order of Ambisonics, it decreases with higher frequencies [9]. The increasing number of Am-

bisonics channels puts up a limit to the method, as both the number of recording microphones and the number of loudspeakers in the array has to be higher than the number of Ambisonics channels. Another issue with HOA recordings is spatial aliasing. Due to the discrete sampling of continuous spatial information, the distance between the used microphones of the array determines a limit frequency above which aliasing occurs. Above this limit frequency, the spatial information of sound events can become ambiguous [29][30].

The method is not bound to microphone recordings of real sound fields, also artificial scenarios can be created and resynthesized for a loudspeaker array. For such artificial scene creation, several methods exist. For example, the toolbox for acoustic scene creation and rendering (TASCAR) enables designing a virtual room, i.e. specifying reflective characteristics of the surfaces in a room. An acoustic model is computed that includes the direct signal path as well as early reflections from each specified surface. TASCAR is capable to be presented on 3rd-Order Ambisonics loudspeaker arrays, though not bound to it [31]. Another approach of artificial scene creation is provided in the *Spat* by *Ircam* toolbox for *Max 8* by *Cycling'74* [32]. Here, the room characteristics can be chosen based on their perceptual influences on the sound sources with no surface characteristics to be specified [33]. The disadvantage of such artificial scene creation methods in comparison to microphone array recordings is that the virtual room has no real reference, while the source room of a recording can be precisely described and explored physically. However, as it has been pointed out in a comparative study by Oreinos and Buchholz [34], the greatest disadvantage of the recording method lies in additional artefacts from the recording itself. In the study, the effect of a complex scene on directional hearing aid algorithms (beamformers) is compared in a real environment, the recording of this environment and the artificial recreation of this environment. Beamformers are phase-sensitive signal processing techniques, therefore are expected to be sensitive to inaccuracy in the virtual scene reproduction method. While all reproduction techniques were concluded to be suitable for beamformer evaluation in hearing aids, some inaccuracy was introduced by the HOA approach. The produced sources led to an increment in perceived width of the target source, while the background noise perceptually did not differ from that of the other methods. It shall be noted that beamformers in hearing aids are considered to be sensitive to phase shifts between the used microphones. Spatial aliasing of HOA towards higher frequencies can limit the applicability of such setups for the evaluation. From the here examined literature, it can be concluded that HOA setups are generally suitable for the assessment of spatial perception with hearing aids, though inaccuracies must be considered depending on the evaluated features [9][30][34][35].

2.3 Sound Quality Assessment

Sound quality is a various term, that includes several different perceptual aspects of a sound event or a sound emitting object. It is a subjective property, which arises from comparisons towards an inner reference that reflects desired features. Good sound quality usually derives from the product to meet the desired features. Sound quality is an affective measure, while the characteristics of a sound event are supposed to be descriptive [36][37]. In sound reproduction industries, improving sound quality means to find out how technical modifications to the product affect the perceived aspects and finally to determine the relationship between these aspects and preference ratings. Besides perceptual evaluation, several methods exist that provide technical measures to describe the quality of signals. For example, total harmonic distortions (THD), the speech intelligibility index (SII) or the Signal-to-Noise-Ratio (SNR) can be considered sound quality measures of signals or sound transmitting objects. However, they do not necessarily correspond to perceived sound quality. This section primarily gives an overview of sensory evaluation techniques of sound quality with a focus on hearing aid evaluation and spatial aspects of sound quality.

In order to measure perceptual aspects of a sensation, the impressions of the subjects have to be quantified. Several methods can be applied in order to perform this quantification. Most methods rely on verbal elicitation, i.e. identifying attributes or labels that best describe the impression. As the perceptual characteristics of sound are various, the general term sound quality is not suitable to summarize all impressions. It is therefore desired to identify the various perceptual dimensions that contribute to the subjective concept of sound quality. The perceptual descriptions of sound can be acquired by letting subjects report their impressions in reaction to a stimulus. These descriptions may differ greatly between individuals, as e.g. their experience or cultural background influence their use of vocabulary. If the impressions of enough subjects are acquired, they can be clustered and filtered for attributes that most of the subjects agree on to be descriptive for the impression. A great deal of effort has been put into the creation of standardized vocabulary or lexicons that characterizes the nature of sound in distinct dimensions. Results are sound quality attribute inventories such as the *Sound Wheel* [37] or the *critical checklist* by the Audio Engineering Society [38]. In such inventories the attribute lists are also grouped into separate categories. It should be kept in mind that not every attribute is suitable to test every product. The researcher has to evaluate which attributes are most suitable to describe the sensations. Additionally to the identification of the attribute labels, each attribute has to be divided into categories that reflect the intensity of its perception. The resulting scales should be designed so that their end labels reflect the extreme limits of evoked perception by the presented material [6].
The development of a standardized lexicon of attributes that describe spatial aspects of sound quality yields additional difficulties. First, as it has been pointed out in subsection 2.1, spatial

hearing in everyday life is a multimodal ability. Secondly, the products that yield spatial information to be evaluated differ greatly in their desired spatial content. Designing a lexicon of suitable attributes to describe the spatial quality of concert hall acoustics, soundbars, headphones or virtual reality applications at once while excluding biases of non-auditory sensations is thus a difficult task. Some lexicons do exist, that aim to describe spatial sound quality, e.g. the extended Soundwheel [39] or the Spatial Audio Quality Inventory (SAQI) [40]. Depending on the product at test, the suitability of the attributes differs. For example, the attribute *Internality* from the Soundwheel corresponds to the impression of sound sources to be located inside the head. Such impressions are not expected when a loudspeaker is evaluated, but becomes important for the evaluation of binaural reproduction systems.

Several methods exist and are well established to perform attribute rating tasks. The International Telecommunication Union (ITU) provides recommendations about sound quality assessment methods. Depending on the products or conditions that should be tested, the methods are more or less suitable. In the case that perceptual differences between products are expected to be large, absolute ratings can be conducted. In absolute rating tasks, each product is presented without a reference. In many cases, the perceptual differences between products in the development state are too small to obtain significantly different absolute ratings. In such cases, comparative ratings are recommended. For only few products to be tested against each other, pairwise comparisons are ideal to check for slight differences. With a larger number of products, such pairwise comparisons become extensive. For conductance of multiple comparative ratings, a standard method is the «multi stimulus test with hidden references and anchors» (MUSHRA). In this method, multiple products are presented on a display, one of which is labelled as a reference condition. The subject is instructed to rate each product against this reference condition on scales that numerically range from 0-100. Included in the list of products is a hidden reference, i.e. the reference condition without being labelled as such. Additionally, anchor conditions are implemented that should reflect the limits of the attribute's scales. The advantage of this method is that small perceptual differences of a great number of products can be assessed at once [6][41]. The biggest disadvantage in such assessment methods is introduced by the use of verbal scales and attribute labels. Rating scales are numerically linear and the attached scale labels are usually equidistant, which implies linearity in the scale labels. This can lead to some bias in the ratings if the assumption of linear scale intervals is not verified. In addition to such biases, the comprehension of the attribute or scale labels themselves can differ between individuals. In a literature review by Zielinski et al. [42], a comparison of results from scaling of labels of an overall sound quality rating across different countries demonstrates differing opinions on the positions of labels along the same scales. For example, the english label «fair» was found out to be considered a more positive label in the USA than in the UK. The issue has to be taken into account especially

when product ratings are translated to be comparable across countries. In order to minimize effects from such biases, it is recommended to conduct such rating tasks with trained assessors. The aim of assessor training is to set a common level of attribute comprehension and a common way of utilizing rating sliders [6].

Because of the potential error sources in verbal elicitation techniques, it is of interest to find non-verbal methods that do not underlie biases by language. There are no standard procedures such as the described MUSHRA-method and in general there are only few studies that conducted such tasks. A method to find out (dis-)similarities between products is called Napping [6]. In a Napping task, subjects are instructed to position the products, representations of the products or product logos on a surface in a way that similarities are indicated by small distances. Non-verbal elicitation techniques gain relevance when it comes to the evaluation of spatial sound quality. A localization task, where subjects point towards the direction of perceived sound sources can be considered a more efficient way to evaluate *localisability* than verbal descriptions of the position. In a review by Mason et al. [43], it was discussed that internal neural processes of non-verbal elicitation methods are closer to the auditory image of space than verbal methods. In an explorative study by Ford et al. [44], subjects were instructed to draw spatial images of a Jazz band presented over loudspeakers on blank pieces of paper. It was concluded that the method is quick and intuitive to be used, containing much information about spatial source-receiver relationship. It was recommended for future studies to provide guide sheets in order to equalize scaling of the auditory event among subjects.

Sound quality assessment with hearing aids underlies special issues in comparison to other electroacoustical reproduction systems. Because the shape and degree of hearing loss differs among hearing impaired persons, the frequency-dependant gain of the hearing aid has to be fitted individually. Additionally, individual hearing loss includes a reduced available dynamic range, i.e. the range of sounds being barely audible and sound that are perceived too loud. The acoustic coupling to the ear causes a loss of natural sound in the non-impaired frequencies. Such individual factors lead to the sound quality preferences in between hearing impaired persons to differ greatly. The highest priority for the application of hearing aids is to compensate for the handicap of impaired speech intelligibility, especially in noisy environments [45].
In the development and provision of hearing aids it has been common praxis to evaluate the quality of hearing aids mostly based on speech intelligibility tests in noise. In the past years, measurements of listening effort has gained attention as it presumably contains more information about perceived hearing aid quality than just speech intelligibility [46]. To investigate the reproduction of auditory space, commonly applied evaluation methods are localization tasks [3]. It was already mentioned that spatial hearing underlies several different aspects. Many features

in hearing aids that are developed in order to facilitate speech intelligibility in complex environments have succeeded in their primary purpose. Though potentially accompanying changes in the spatial impression of the hearing aid wearers has barely been investigated.

In the MarkeTrak VIII survey by S. Kochkin [7], the top 10 factors that contribute to overall hearing aid satisfaction are listed. Five of these factors can be related to descriptions of sound quality. Sound quality assessment with hearing aids become more important in the context of audio streaming features.

The great differences in preference among hearing aid users produces another issue when it comes to sound quality assessment. Such surveys can thus be performed with trained normal hearing subjects that rate the quality of simulated hearing aid algorithms or pre-recorded sound samples. It is expected that these assessor groups are sensible to even minor sound quality impairment. Though it has to be kept in mind, that generalizations of the outcomes on hearing aid users are limited. Consequently, it is of interest to employ a homogeneous group of hearing impaired users for sound quality evaluation [6].

3 Materials and Methods

3.1 Acoustic Stimuli

3.1.1 Scene Design

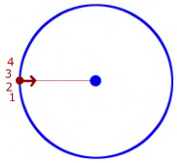
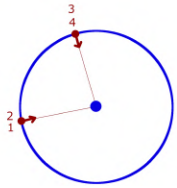
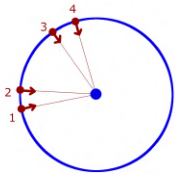
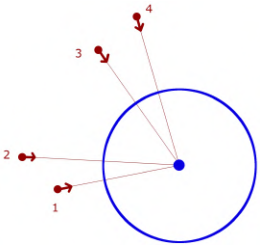
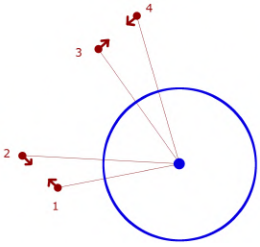
As described before, spatial hearing is a multimodal ability, meaning that more than just the auditory input contributes to the acoustic impression of space. The most dominant cue for orientation in space is provided by the visual input. Consequently, in order to specifically measure auditory spatial perception, the visual domain has to be excluded from such experiments. This constitutes a large disadvantage when it comes to the creation of ecologically valid listening environments. Accepting the fact that it is close to impossible to create realistic listening scenarios while withholding visual information led to the realization that no reported room impression can be considered comparable to that of a real situation. This resulted in the concept of designing listening scenarios that deliberately evoke different degrees of naturalness that shall be rated relatively to one another.

In this work, a set of five scenes is created that only differ in their spatial information (see Table 3.1). A scene consists of four speakers that form two independent dialogues that are looped at one minute. The concept of inter-scene variations is to go from least (Scene #1) to most ecologically valid scene (Scene #5) in discrete steps. Speaker alignments such as in scene 5 can often be found in real life, e.g. in an office- or canteen-situation, where the listener is surrounded by two dialogues. The listener can be considered uninvolved, as she/he is not addressed by the content of the dialogues. However, in scenes #1 to #4 the perceived involvement of the speaker might change due to the single speakers' facing angles. In these scenes the listener might experience contradictory perception, as the acoustic information indicates being the addressed person while the content of the dialogues indicates that he is not.

In order to provide as much spatial information as possible while still excluding visual input, it would be ideal to recreate the scenes with loudspeakers in a real room and a curtain around the subject. However, this method could not be set up at reasonable expense, also because changes of scenes should be executed as fast as possible. The alternative method of choice in this work is using HOA playback. Part of this experiment is to validate the method of HOA together with convolution of dry sources to create more or less realistic scenes. This method comes with several advantages in this context. For one, the subjects are allowed to turn their heads within the sweet spot, which enables making use of sensory-motor feedback as an important spatial cue. Also, a change of scenes can be done almost instantly which includes a change of speaker alignments or even the room where the scenes take place. Another advantage is that once generated, the HOA-room impulse responses can be convolved with different stimuli and mixed to any desired extent.

To make use of the advantages of HOA-playback, all scenes have been recorded in two rooms.

Table 3.1 – Display of the created scenes and their description. The blue dot indicates the listening position; the blue circle indicates the loudspeaker ring around the listener at 1.5 m. The red dots each represent a physical sound source. The numbers indicate the single speakers. Speakers 1 & 2 and speakers 3 & 4 each form a dialogue.

Scene #	Visual Representation	Description
Scene 1		• Least ecologically valid scene of the set • All speakers sound from one source
Scene 2		• The two dialogues are laterally split • Each pair of speakers sound from one source
Scene 3		• The speakers are laterally split from their conversing partners • The lateral speaker positions are equal to those in scene #2
Scene 4		• All speakers are moved to different distances • The distance between two conversing partners is set to 1 m
Scene 5		• All speakers are facing towards their dialogue partners • Most ecologically valid scene of the set

The choice of rooms is expected to have a high impact on most aspects of spatial perception. In this work, two rooms from within Sonova headquarters in Stäfa, Switzerland have been chosen. The rooms are referred to as Gravel (room G) and Kircher (room K). The schematics of both rooms with the speaker alignments of scene #5 are displayed in Figure 3.1 and corresponding pictures in Figure 3.2. Both rooms are used as meeting rooms and have acoustically similar designs, meaning the same carpet (on the floor and on some walls), the same ceiling and the same height. The base area of room G can be considered approximately twice as large as in room K, though with a slightly different layout. The ceiling of room G has an overhang close to the recording position which is lower than the rest of the ceiling. It is worth mentioning that the ceilings in both rooms are wave-shaped (see Figure 3.2). This ceiling design has the function of selective diffusion of a speaker's voice similarly to that of a pulpit. For practical reasons, the tables in both rooms have not been removed during the Room Impulse Response (RIR)-recordings, which is a potential source of noise in the perception of the sources. All tables are equipped with sockets and network adapters, thus fixed to the floor. The speaker positions have been chosen such that in room K, the positions in the more distant scenes (#4 and #5) take the room limits into account. Transferring the same scenes from room K to G with the exact same speaker positions caused some speakers to be placed on the tables. The microphone height has been set to 1.2 m, as this equals the sweet spot height in the testing chamber. The height of the loudspeaker has been set to 1.5 m to the ground, which is thought of as a mean height between standing and sitting position. The listening position in both rooms has been chosen relatively to the two adjacent walls (1.1 m and 1.9 m). The speaker positions are relative to the listening position. The facing angles of the loudspeakers has been changed by turning them on their stands, thus ignoring the acoustic centre of the speaker. This has been done because a person turning their head also does not turn around their acoustic centre but around their spine.

3.1.2 Room Impulse Response Acquisition

The RIRs have been generated utilizing the exponential swept-sine method [47]. This technique is based on the convolution of an exponential sine sweep and its inverse filter. The result of such a convolution is a delta impulse. If the sine sweep is sent through a linear time-invariant (LTI) system, the impulse is formed by the system's response. A room can be considered an LTI-system. The RIR is generated by recording the exponential sine sweep in the room and convolving it with its inverse filter. In order to enhance the SNR of the RIRs, five sine sweeps are recorded and their average will be convolved with the inverse sweep. The averaging method is not robust against even slight time invariances, which are not expected to come from a room. The generated sweeps are 5 s long with equally long pauses in between. The RIR is specific for the room characteristics and the relationship of positions of speakers and listener. Consequently,

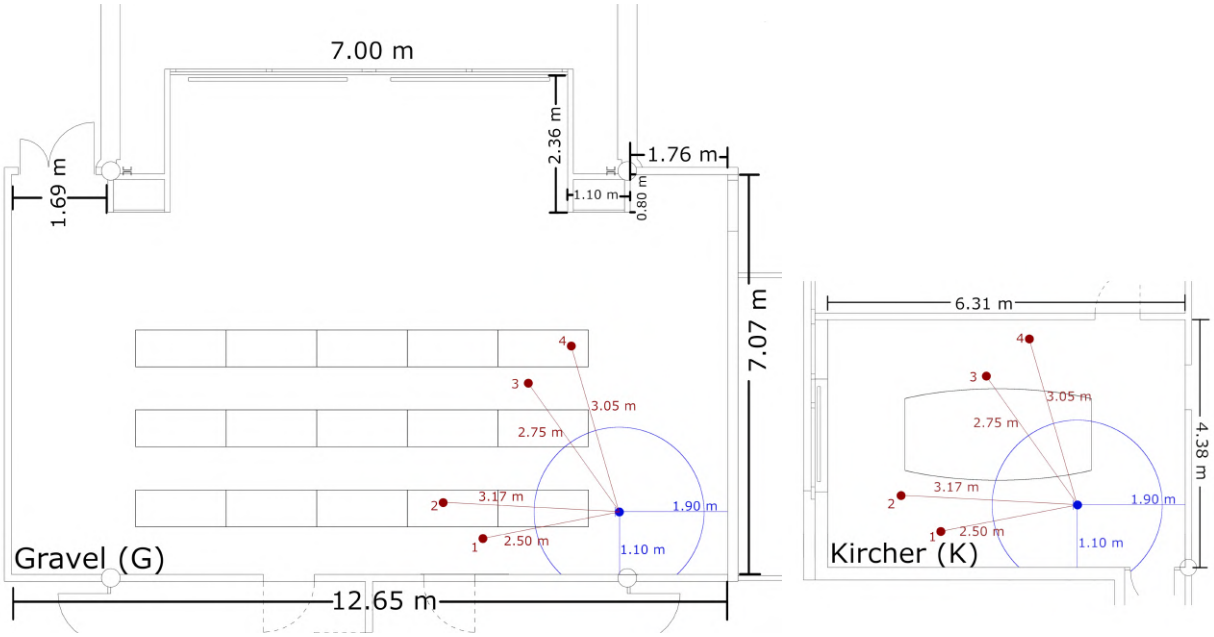


Figure 3.1 – Display of scene 5 in both rooms Gravel (left) and Kircher (right) using same scale in a perspective from above. The red dots each represent one speaker. The blue markings indicate the listening position. The distances to the adjacent walls are equal in both rooms. Also displayed are the tables as obstacles that have not been removed during recording in both rooms.

Figure 3.2 – Pictures of both rooms Gravel (left) and Kircher (right) during the recordings of the room impulse responses. The recording equipment (Eigenmike, loudspeaker, laptops and sound cards) can be seen in both pictures.

for the generation of the set of scenes, each loudspeaker position in each room including facing angle changes requires one RIR.
In order to generate HOA scenes, the RIRs have been recorded with the *em32 Eigenmike* by *mh acoustics*. The *Eigenmike* is an array of 32 spherically aligned microphones, capable of fourth order Ambisonics output. The recording setup can be seen in Figure 3.2. The *Eigenmike* comes with a dedicated sound card and the *Eigenstudio* driver software. Recordings made with this setup are automatically encoded in the 25 channels of the 4th Order Ambisonics format. The raw 32 channel recordings are also saved in case of encoding strategy revision. The whole RIR-generation procedure is done with each of the 25 Ambisonics channels separately. As mentioned in Section 2.2, HOA recording produces an upper limit frequency, above which spatial information becomes inaccurate. According to the release notes of the *Eigenmike*, this frequency is at 9 kHz [48]. According to a project thesis from the institute for electronic music and acoustics IEM in Graz, Austra, the limit frequency has been found to be 5.2 kHz [49]. Consequently, the inaccuracy for higher frequencies has to be taken into account depending on the products at test. The exponential sine sweeps are generated with the chirp function in *Matlab 2017b*. The sine sweep playback has been done in *Adobe Audition 3.0* with an *RME Fireface UC* and a *Genelec 8020* loudspeaker. Because the goal is to reproduce human speech through the loudspeakers, it would be ideal to generate the sine sweeps from a human head-shaped speaker. However, these are hard to acquire and their output power is spectrally limited especially when it comes to more distant positions. The Genelec speaker has been chosen based on their head-like size and it's rounded edges. Prior to recording, it has been made sure that facing angle changes of the speakers are perceived by doing informal blindfolded surveys with expert listeners from Sonova. The Sound Pressure Level (SPL) of the sine-sweeps has been set once for all recordings to 70 dB(A,eq) of the first scene's loudspeaker position. The clocks of the two sound cards (one for recording, one for playback) have been synchronized to avoid any jitter issues.
The RIR decomposition of the recordings is processed in Matlab. To reduce the enormous processing time because of the multitude of recordings (25 channels * 13 positions * 2 rooms), the actual convolution is performed by calculating the products of the zero-padded signals in the frequency domain. Exemplary RIRs for both rooms are displayed in Figure 3.3. It can be noted that the RIRs show a high SNR. The reverberation times RT60 can be computed from the RIRs and are displayed frequency dependent in Figure 3.4.

The actual scenes are generated by convolving sound files with the RIRs. The sound files in this study are supposed to be dialogue files. The demand on these sound files was to have two German speakers that are separated on two single channels with desirably low noise. The content of the dialogue is not of high interest, it should primarily sound natural in the context of a dialogue in person (e.g. a telephone conversation could be audibly identified as such by the content). Also, there should be no influences by the microphone (e.g. no breathing noises due

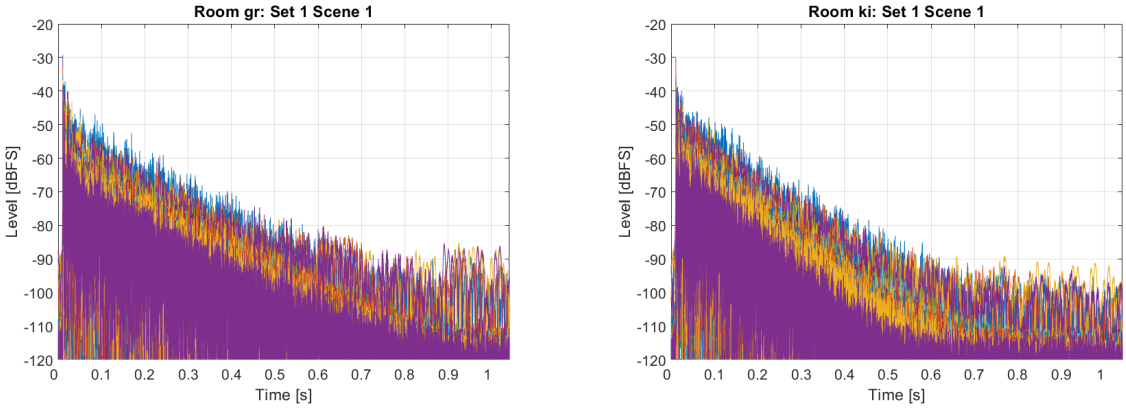


Figure 3.3 – Display of the impulse responses in dBFS over time in s for both room Gravel (left) and room Kircher (right) of scene #1. The impulse responses of each Ambisonics channel are coded with a different colour.

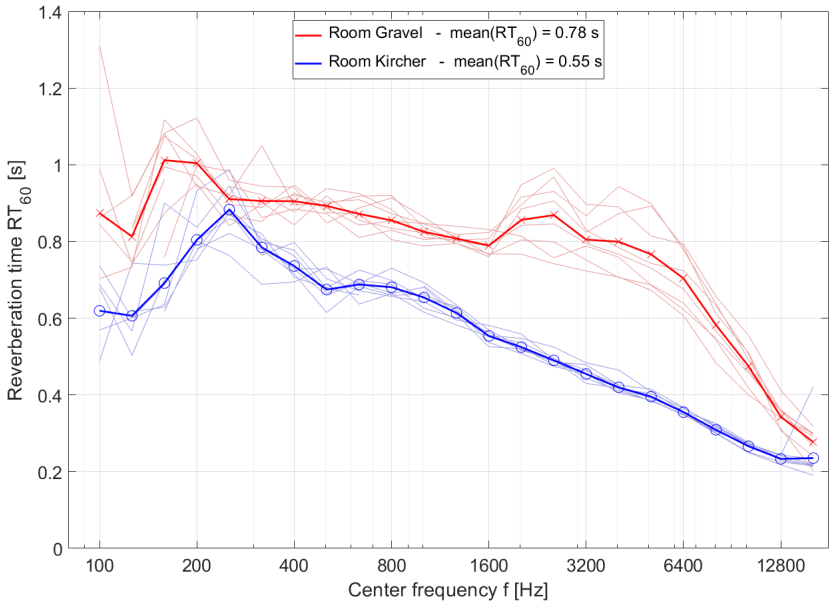


Figure 3.4 – Display of the frequency-dependent reverberation time RT60 in third octave steps. The light lines represent the RT60 curve for each speaker position. The thick lines represent the mean RT60 of each room. The red line indicates room Gravel, the blue line indicates room Kircher. The mean RT60s of both rooms is displayed in the legend.

to misplaced headset microphone), the active speaker should change quite frequently and some diversity in the voices is desirable. It should sound as natural as possible, the speakers should sound like they could really be in the same room as the listener.

One of two dialogues has been found in the library of Skowronek et al. [50]. The library contains multi-channel conversation files of up to six speakers that has been produced as test signals for quality assessment of audio conferencing systems. Scene 13 from the third DVD of the library contains a dialogue of two male speakers that match most of the requirements. The voices are

well distinguishable, there are no breathing noises or statements about transmission quality or similar. The dialogue is about finding a suitable area for the location of a movie set, while one speaker is the person describing the demands on the location and the other agrees to search for the location. Speaker one is talking 58 % and speaker two is talking 38 % of the time.
The other dialogue has been taken from a sound DVD that is provided with the *Phonak Target Fitting Software*. The DVD contains several sound samples that are provided for hearing aid fitting audiologists to provide some natural sound samples for customers. The chosen dialogue is spoken by two professional speakers, one male, one female. The dialogue exists in three different versions, that differ in the level of noise that has been played on the headphones of the speakers in order to include the change in vocal effort with varying environmental level (Lombard Effect)[51]. In this work, the dialogue of medium noise interference has been chosen, as it sounds natural considering the presence of two interfering speakers in the same room. The dialogue is about the male speaker inquiring about work related issues of the female speaker. The female speaker is talking 57 % and the male speaker 28 % of the time.
Each speaker is now assigned a position in the scene layouts (see Table 3.1). The two speakers from the first described dialogue are assigned positions 1 and 2, and the speakers from the second dialogue are assigned positions 3 (female) and 4 (male). Each speaker's channel is now convolved with the corresponding RIR, afterwards the scene is created by simply adding each single speaker's convolved stream. For scene #1, all four speakers share the same RIR, for scene #2, the dialogues each share one RIR. It is worth mentioning that once acquired, the RIRs can be used and combined in different ways. If a scene of four simultaneously active speakers forms too demanding tasks, the scenes can easily be reduced to simpler layouts.
In Figure 3.5, the effect of the facing angle and distance change is displayed. The dashed lines mark the RMS-level of scene #3, the upper area limits mark that of scene #4 and the lower limits mark that of scene #5 for each speaker and both rooms (red: Gravel ; blue: Kircher). The RMS levels are calculated from the convolved sound files of the first encoded Ambisonics channel, i.e. the channel that contains the omnidirectional information. The effect of the distance change can be seen as a broadband shift. The effect of the facing angle change can be seen as a low pass filter. The intensity of the effects is dependant on the speaker and the room. The facing angles are always changed so that the dialogue partners face each other. Consequently, one of both speakers is always facing farther away than its partner, except both have the same distances to the listener. For example, as speaker 1 is turned farther away than his conversing partner, the low pass filter effect is greater. It is expected, that the accuracy in facing angle or distance perception differs throughout the single speakers.

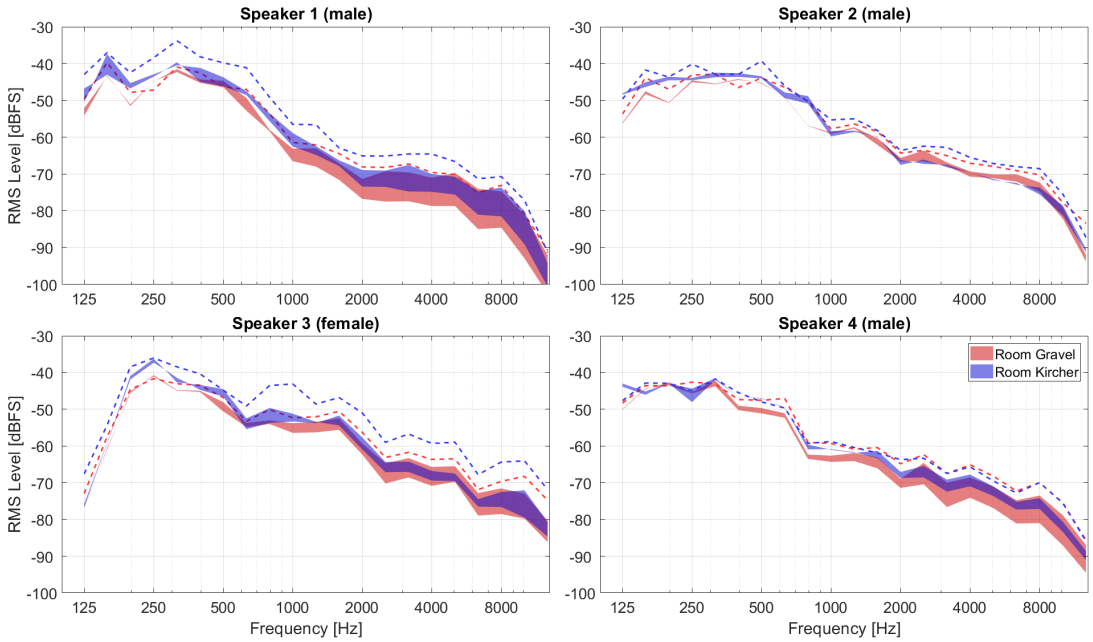


Figure 3.5 – Display of the frequency-dependent effect of facing angle and distance change on the RMS level for the four speakers in both rooms (room Gravel and room Kircher). The dashed lines mark the RMS level of scenes #3 (speaker close, facing listener), the upper area limits mark that of scene #4 (distance change) and the lower limits mark that of scene #5 (Speaker facing its conversing partner).

3.1.3 Scene Playback

The experiment took place in a non-anechoic room with a low reverberation time ($RT_{60} = 199\,ms$). The room is equipped with a spherical array of 32 *MeyerSound MM-4XP* loudspeakers and two *MeyerSound MM-10XP* subwoofers (see Figure 3.6). The computer and the audio interfaces are placed outside of the test chamber. The loudspeakers inside are aligned as follows:

- eight loudspeakers at $-30\,^\circ$ elevation
- twelve loudspeakers at $0\,^\circ$ elevation
- eight loudspeakers at $+30\,^\circ$ elevation and
- four loudspeakers at $+60\,^\circ$ elevation.

All loudspeakers are pointing towards the centre of the sphere and are placed equispaced in their horizontal plane. The distance from the sphere's centre to the loudspeakers is $1\,$m for the top speakers and $1.5\,$m for all the others. The loudspeakers are covered by a circular two layer fabric curtain which can be considered acoustically transparent. The digital signals are processed by a *RME HDSPe FX* soundcard, DA-converted by a *DirectOut Andiamo 2.DA* converter and pre-amplified with *MeyerSound MPS-488P* power supplies. The two subwoofers are placed on the sides of the room. The signals from all other speakers are routed to the subwoofers which mainly

affects the spectral range below 200 Hz. The SPL of the subwoofers is adjusted so that the scene perception is subjectively natural.

The playback of the generated HOA-scenes is performed with the toolbox *Spat* by *Ircam* [33]

Figure 3.6 – Picture of the testing chamber. The loudspeakers are aligned spherically around the centre of the room. The curtain is acoustically transparent and shall avoid visual bias by the true loudspeaker positions or the true room dimensions. The 'X' on the floor marks the position of the sweet spot.

for the visual programming language *Max 8* by *Cycling'74* [32]. The spat toolbox is capable of decoding and playback of HOA-signals. Input parameters for the toolbox are the speaker positions (azimuth and elevation angles) and the type of ambisonics decoding strategy. In this experiment, an energy-preserving decoder is chosen, as it is robust against non-uniformly distributed loudspeaker arrays [52].

A program in Max 8 is written that manages scene playback, changes in the signals (change of room or scene) and a loop function. The demand on the program was that each change or restart of the material should run smoothly, meaning with no audible transition noises. When the scene is changed, the next one should continue playing at the same time. As a safety measure against buffer dropouts due to the large file size of the scenes, the scene playback is implemented with two player elements with a crossfader in between. When the next scene is started, the inactive player gets the next file and the gains of both players are faded in between. As a consequence, the transition time between the scene changes is now $2\,s$. Besides being a working safety measure against buffer dropouts, the transition noises are also reduced to a minimum.

The experimental design requires it to steer scene playback remotely from within the test chamber. The actual test programs (Section 3.2) are all run on Matlab 2017b on an additional laptop. The remote communication with the HOA-scene player (selecting the scene /room; pause /resume playback) has been realized using User Datagram Protocol (UDP) over the network. The UDP-messages are sent from Matlab on the laptop to Max 8 without any perceivable delay. The disadvantage of UDP is that there is no feedback of successful transmission. As a safety measure, a wired connection of the laptop with the network is established. No failed transmissions have been noted during the experiment.

3.2 Experimental Design

3.2.1 Subjects

The goal of this experiment is to find out if the designed scenes and tasks enable differentiable assessment of perceived spatial sound quality. In order to optimally rate the effect of the scenes or tasks on the perception, it is desirable to minimize effects by the subjects or the replication of the experiment. Consequently, the subjects have been selected from an expert listener panel of Sonova employees. Expert listeners are trained assessors whose qualities are rated based on the follwing three performance characteristics:

Repeatability:	The assessor's test ratings show little variance compared to the retest ratings.
Discrimination:	The assessor sensitively rates perceptual differences on attribute scales.
Agreement:	The assessor's ratings show little variance compared to the panel group.

The expert listener training has been designed and conducted by *DELTA SenseLab* which was focused on the rating of hearing aid related sound quality attributes. It is worth mentioning that the expert listeners are not specifically trained on rating spatial sound quality.
From the expert listener panel, 11 subjects (six female, five male; age ranging from 28 to 41, mean 34.4) of German or Swiss-German mother tongue were invited to the experiments. The experiment was done in two appointments, while the second one took place one week after the first and served as retest appointment. The normal hearing ability of the expert listener panel is controlled on a regular basis. Four tasks have been developed, which are described in the following sections.

3.2.2 Task 1: Familiarization Phase

The familiarization phase was only done at the first appointment prior to the other tasks. The idea was primarily to let the subjects experience all the parameters that can change throughout scene or room variations themselves. The subjects should be sensitized to those aspects in the scenes that change and to those that do not. All the created scenes have been played in a fixed order, which is displayed in the schematic of Figure 3.7. The order of the scenes is chosen to give the subjects more and more spatial information by each step. A scene change has been initiated by the investigator who sits right out of the loudspeaker array. After each change of scenes, the subject is asked "Did you perceive a change in the scene? How would you describe it?". The answers are openly logged by the investigator in a Graphical User Interface (GUI) in Matlab (see Figure 3.8), that also sends the scene change messages to Max 8. The subjects were allowed to listen into the previous scene again. It can be expected that the subjects report their impressions without any bias by giving them attributes to describe them or any hints of what might change.
In the later tasks of the experiments, the scene order is randomized and all scenes are compared relatively to all others. This might result in smaller perceptual differences e.g. in between scenes

#2 and #3 not to be reflected in the outcomes at all. The logged responses from this task could serve as proof if the subjects were able to hear such smaller differences.

The results will be classified if the subjects report the perceptual changes that were intended

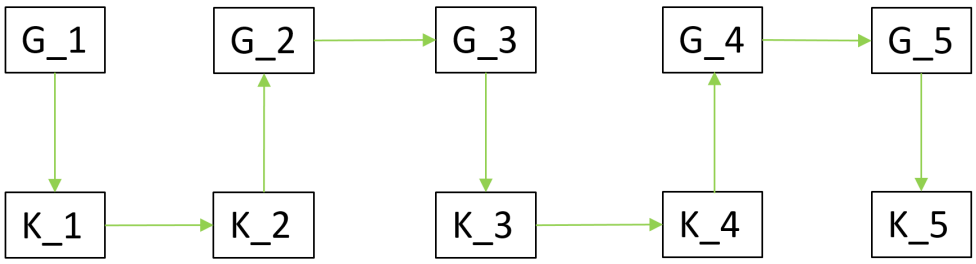


Figure 3.7 – Schematic of the procedure in the familiarization phase of the experiment. G and K stand for rooms Gravel and Kircher, the numbers indicate scene indices. The green arrows show the fixed order of scene changes in this phase.

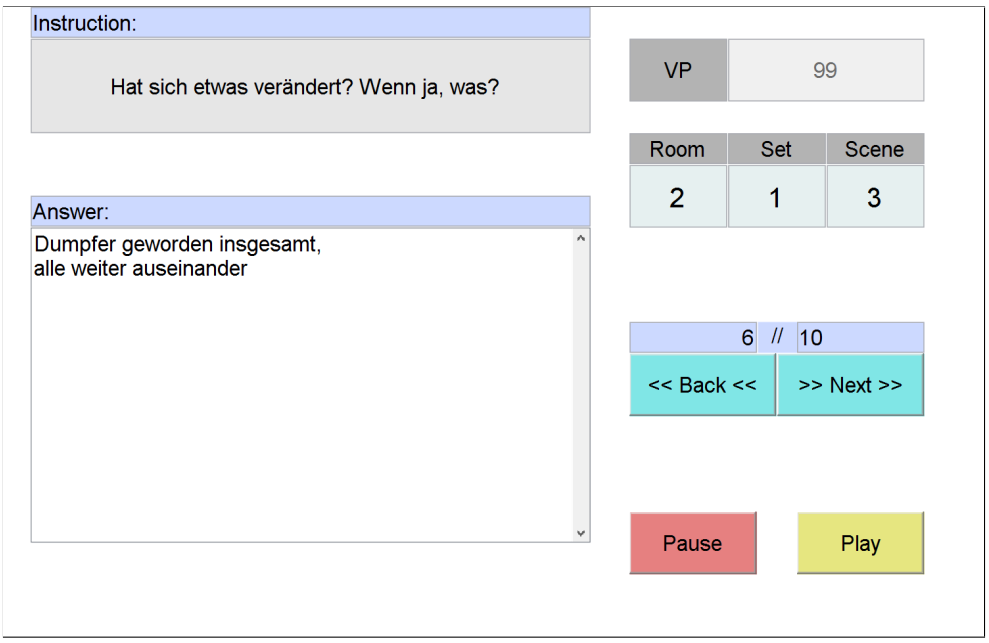


Figure 3.8 – Graphical User Interface in Matlab for the familiarization phase. The instruction reads «Did anything change? If yes, describe the change.». The answers are logged in the text box, the example reads that it now sounds muffled and that all speakers seem to be further away from each other. The Next and Back buttons enable comparisons with the previous scene.

during scene design. It is expected that each change of rooms is responded with a description of room size or reverberance. Scene variations 1 to 2, 2 to 3 and 3 to 4 are mainly expected to be perceived as changes in the speaker positions laterally or in the distance. It will be specifically interesting to see how the scene change 4 to 5 will be responded with, as subjective tests prior to the experiments revealed some difficulties in distinguishing distance from facing angle variations. Also it can be assumed that the expert listeners are not used to listen to sound sources that

are facing away, despite this being a natural scene layout.

3.2.3 Task 2: Attribute Rating

This section addresses the second task of the experiment, which has been conducted at both appointments. The scenes were designed to evoke distinct spatial impressions, thus the assessment tasks should reflect all perceptual changes in the scenes equally distinct. Attribute rating tasks are established methods when it comes to sound quality assessment [6]. Before conducting the actual rating task, a list of attributes that reflect the scene variations has to be defined.

Attribute Selection The strategy to find a list of suitable attributes for this task was to put an extensive list of attributes up to discussion with a few expert assessors, thus not to be confused with the subjects from the expert listener panel. Five such expert assessors are chosen based on their specific expertises. These were ranging from domains of physical spatial acoustics, virtual acoustics or subjective sound quality assessment. Because of their expertises and partly their involvement in the concept of this work or similar studies, two non-expert assessors that were completely naive to the subject have been included.
The preselection of attributes was performed similar to the familiarization phase of the experiment. The expert assessors were offered a GUI which allowed them to switch scenes and rooms, while they were asked to verbally describe all impressions. The assessors were also specifically asked for labels to their impressions. In addition to the subjective descriptions of the scenes, some attributes that are found in the literature (e.g. Soundwheel [39] and SAQI [40]) and match the concepts of the scenes are put up for discussion. The demand on the discussion's outcome was to find attribute and scale labels that naive subjects all comprehend similarly and that they are suitable to describe the presented scenes. In Table 3.2, an overview of the final list of attributes is shown. The subjects were all native German or Swiss-German, so the attribute and scale labels have been presented and developed in German. This list has been tested on the two naive listeners and has found approval. It is worth mentioning, that this is no proof of a fitting list of attributes regarding the small sample size of assessors and their individual backgrounds. The results of this task in combination with those of the familiarization phase will indicate how well the attributes meet the experimental conditions.

Table 3.2 – English translation of the final list of attributes for the comparative rating task along with the instructions and scale end labels. The subjects were displayed German attribute labels, instructions and scales (see below).

Label	Instruction	Scale end labels
Distance	How far do you perceive yourself away from the scene?	very close - very distant
Reverberance	How reverberant does the scene sound?	dry - reverberant
Listening Effort	How difficult do you find following the speakers?	effortless - extremely effortful
Room Size	How big do you perceive the room of the scenes?	very small - very large
Realism	How realistic does the scene sound?	very unrealistic - very realistic
Distributedness	How much are the speakers distributed in the room?	not distributed - very much distributed

Table 3.3 – Original German list of attributes for the comparative rating task along with the instructions and scale end labels.

Label	Instruction	Scale end labels
Distanz	Wie weit bist du von der Szene entfernt?	sehr nah - sehr fern
Halligkeit	Wie hallig nimmst du die Szene wahr?	trocken - hallig
Höranstrengung	Wie anstrengend ist es, den Sprechern zu folgen?	mühelos - extrem anstrengend
Raumgrösse	Wie gross nimmst du den Raum wahr, in dem die Szene stattfindet?	very small - sehr gross
Realismus	Wie realistisch klingt die Szene?	sehr unrealistisch - sehr realistisch
Verteiltheit	Wie stark sind die Sprecher im raum verteilt?	sehr wenig verteilt - sehr stark verteilt

The discussions have shown that some attributes are difficult to describe the whole scene as one, as there might be differences in between single speakers. Consequently, the list has been made of attributes that are all applicable to the whole scene. Furthermore, it has been emphasized in the instructions to the subjects that they have to rate the whole scene and always consider all four speakers. Some considerations about the attributes are listed below:

Distance is an example for an attribute that is applicable for the single speakers rather than the whole scene. Technically, there are only two levels of distance that can be expected (in

scenes 1 to 3 and scenes 4 to 5).

Reverberance is expected to be rated differently in reaction to changes of distance, facing angle and room.

Room Size together with Distance and Reverberance produced the best consensus among all expert assessors. As there are only two rooms and thus room sizes at stake, it will be interesting to see if the ratings reflect more than two levels.

Listening Effort was not literally described by the assessors but matches frequent descriptions such as «It is hard to listen to them when they are further away». The label *Listening Effort* in this case is not only referring to the ease of following speech but also reflects the ease of distinguishing the single speakers spatially. It is expected to be rated high for scenes #1 and #2, where more than one speaker share one source position and highest when all speakers are turned away in the distance.

Realism was the attribute that was most frequently discussed about. For one, it can be noted that each assessor seemed to have a different opinion on realism in this context. Partly it was described as unrealistic to perceive reverberation in an obviously dry room, partly it was perceived unrealistic when all speakers are facing the listener. The descriptions and the inner reference also seemed to depend strongly on the amount of background knowledge, which differed greatly between the expert assessors. It is assumed that, if properly introduced, there might be more consensus on the attribute in between the naive subjects. The attribute has not been rejected for the list despite the poor consensus among expert assessors. It is believed to give information about the quality of the HOA system and to reflect the concept of scene variations, going from least to most ecological validity. In order to get some equality in the comprehension of this attribute, the subjects are instructed to ask themselves, if they could imagine the scene to take place as perceived.

Distributedness, despite being a made up word, found high consensus among expert assessors. It shall simply describe the positions of all speakers relatively to another and to the listener. It matched verbal descriptions like «Now it sounds like they moved away from one another». It is expected to be rated low for scenes #1 and #2 and high for scenes #4 and #5.

Comparative Rating For a multiple comparisons rating task, a GUI is provided to the subjects in Matlab. The subjects have the laptop with the program on their laps while situated in the sweet spot. A display of the used GUI at the example of the attribute *Distance* is shown in Figure 3.9. The user can initiate the playback of a scene by pressing on one of the buttons A-G. After a delay of two seconds due to the crossfading (see Section 3.1.3), the corresponding slider

gets active. The sliders are limited from 0 to 100, with a minimal step size of 1. In each window, seven scenes are presented. Additionally to the five scenes of one room, scenes #1 and #5 of the other room are implemented so the ratings of the two rooms are comparable. The order of the seven scenes is randomized. When all scenes of one trial are rated, the button «Weiter» (english: Next) gets active and initiates the second trial, i.e. the set of scenes from the other room. The order of rooms and of the scenes was randomized for each attribute and each subject. Including scenes #1 and #5 of the other room was done in order to set the upper and lower limits of the rating scales. That makes the ratings of the two rooms comparable. However, this is based on the assumption that no other scene than scenes #1 and #5 are rated lowest or highest.

There are many well established methods of conducting such multiple comparisons rating tasks.

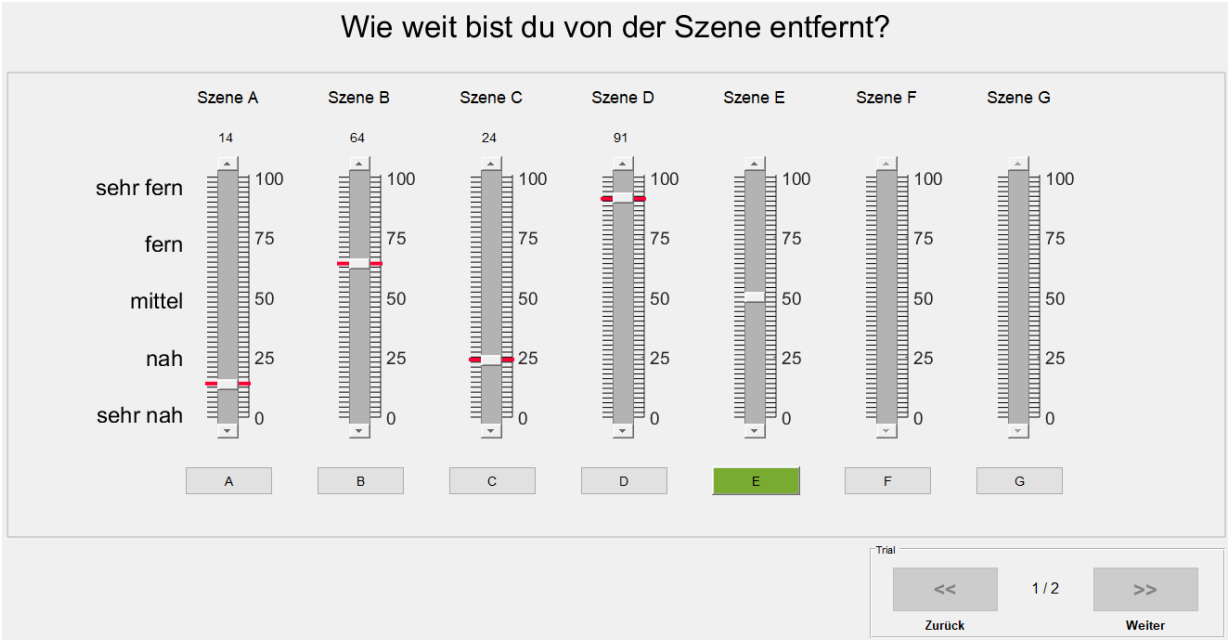


Figure 3.9 – Graphical User Interface of the comparative rating task for the attribute *Distance*. A press on one of the buttons A-G initiates a scene to be played back. The green mark indicates which scene is currently playing. Each scene has to be rated by actuating the corresponding sliders. A slider gets active after a scene button has been pressed. The button «Weiter» (english: Next) initiates the next set of seven scenes and becomes active after all sliders have been activated. The heading translates to «How far do you perceive yourself away from the scene?»

Commonly used are approaches that make use of (hidden) reference and anchor conditions such as the «Multi Stimulus Test With Hidden Reference and Anchor» (MUSHRA) [41]. In this context, defining a reference or anchor condition would mean to make assumptions of how all other scenes are rated, as all scenes would have been rated relatively to this reference. This might distort the ratings of a set of scenes. Because the ratings of these conditions cannot be assumed to be absolute, they would need to be excluded from the analysis of results. Due to the distortion in the ratings or loss of information, it has been decided not to implement such mechanisms. This raises the effort for the actual task, as the scenes cannot be rated in a pairwise comparison (refer-

ence to A/.../G) but have to be rated relatively with all other scenes. However, the expert listeners should be capable of such multiple comparative ratings due to their training. There is no strategy specified how to do such rating tasks. A common approach was to make a first intuitive rating of all scenes based on the scale labels and then sort them to rate the small perceptual differences. Also, the expert listeners are trained to make use of the whole presented scale.
At this point of the experiment, it is revealed to the subjects, that three parameters are changed throughout scene variations: Speaker positions, their facing angles and the room where the scene takes place.

3.2.4 Task 3: Figurine Alignment

The implementation of the attribute rating task points out some issues regarding the assessment of such complex scenes. First, the attributes might cause confusion because they are not easily applicable to the four speakers simultaneously. Second, all subjects need to have a similar understanding of the attributes, their instructions and their scale labels. This might be less of a problem when the assessors are somewhat trained and involved with audio topics. But when it comes to assessing the spatial perception of hearing impaired or hearing aid users, more noise can be assumed in the ratings due to differing comprehension of the attributes. The figurine alignment task has been developed to provide an intuitive task, that subjects have the same comprehension of irrespective of their background.
The task for the subjects is to align four figurines on a grid just like they perceive the speakers in the scenes. In Figure 3.10, the requisites for the tasks are shown. Fourteen scenes are presented, that are compiled like in one attribute rating window in the previous task (two rooms, five scenes and two each of the other room). The scene playback is initiated by the investigator who is situated right out of the loudspeaker array for the task. When the recreation of a scene is completed by the subject, the investigator is called who then initiates a snapshot with a webcam above the grid. To compensate for the limited resolution of the webcam, the figurines were marked with reflective stickers on their heads. To facilitate detecting the facing angle of the figurines, the subjects were instructed to keep the arms of the figurines raised.
There is one considerable flaw in the design of this test. In order to reposition the figurines, the subjects have to bend over to reach the carpet. Consequently, with each figurine alignment, they bend out of the sweet spot, thus distorting the spatial perception of the scene. The subjects are instructed to ignore any differing impressions when realigning and to only listen to the scene in an upright position. It has also been taken care of only recruiting subjects that are well capable of bending over from the chair.
The decision of conducting the task on the carpet was primarily made so that there are no HOA-influencing obstacles and because of an easier size ratio scaling for the grid. The grid contains

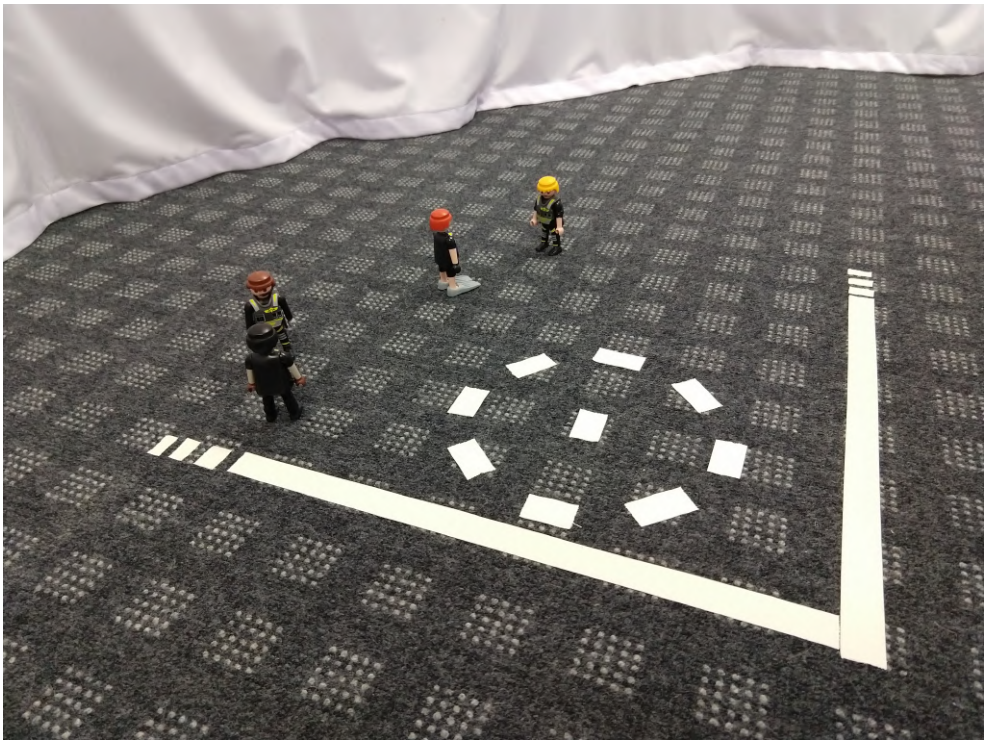


Figure 3.10 – Picture showing the used figurines on the grid for the scene recreation task. The grid was marked on the carpet in front of the subject's position on the floor.

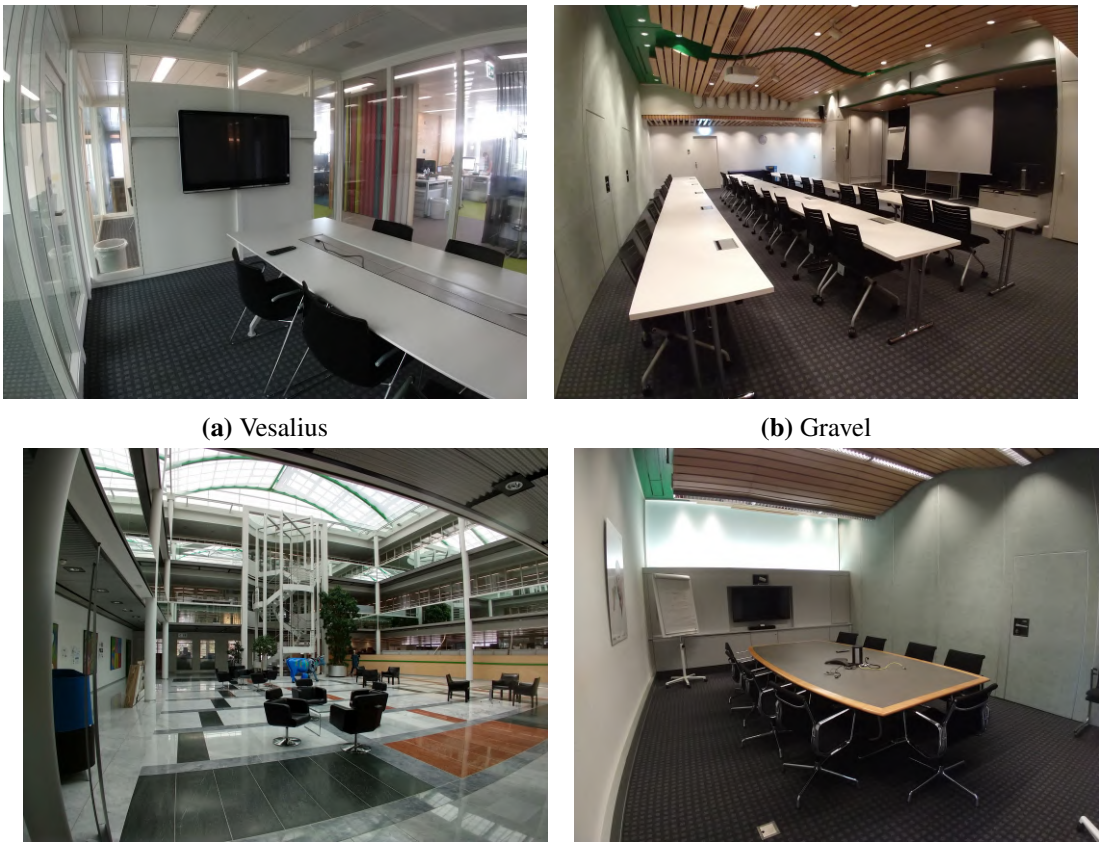
a dotted circle, which indicates the curtain with a dot in the middle as the listener. The two adjacent walls are marked with the solid lines, whose dotted ends indicate that the other two walls are not defined. The distance of the curtain to the subject is chosen to be the same as the height of a figurine's chest (real curtain distance: $1.2\,\mathrm{m}$).
The outcome of this task shall provide individual information that cannot be assessed with the attribute rating task. It is expected that the scene recreations reflect the perception of each speaker individually. It is emphasized to the subjects that the facing angles of the figurines will be analysed as well, which is an aspect that is not easily assessed in an attribute rating task. A disadvantage of this task is that scenes #1 and #2 are physically not recreateable with the figurines. In case of the subjects asking for this problem (which happened in two cases), they are instructed to position the figurines as close as possible.

3.2.5 Task 4: Room Characterization

The room characterization task was conducted only in the second appointment as the last task. The figurine alignment task provides information about the single speakers of a scene, though no direct information about room perception. This task has been designed in order to get as much information about the imagined room where the scene takes place. Prior to the task, it was revealed that there are only two different rooms.

Scenes #5 of both rooms have been presented by the investigator once more. Scenes #5 were chosen because their signals are influenced by the room responses the most. The subjects are instructed to give subjective feedback of how they imagine the rooms where the scenes take place. It was asked for floor and wall materials, if there are many obstacles inside, what purpose the room might serve and if they have a real room in mind. The answers were openly logged by the investigator.

As a second part of this test, pictures of four rooms (Figure 3.11) are layed out in front of the subjects. Pictures **(b)** and **(d)** show the real rooms. The subjects are instructed to assign the heard scenes to one of the pictures, knowing that the true rooms are included. All displayed rooms are known to the subjects, as they are within the Sonova building in Stäfa, Switzerland. From the first impressions of the scene playback in the testing chamber and the expert feedback during attribution phase, it is known that the scenes are generally perceived more reverberant than the real rooms. This task is designed to get some evidence for this impression. The pictures are chosen based on their 'visual reverberance', meaning that they contain reflective material. It is expected that subjects will tend to overestimate the size and reverberance of the rooms.



(a) Vesalius

(b) Gravel

(c) Quattro

(d) Kircher

Figure 3.11 – Display of the four rooms that were presented in the forced choice task. The labels refer to the names of the rooms. The true choices are: **(b)** as room Gravel and **(d)** as room Kircher.

4 Results

4.1 Familiarization Phase

The outcome of the familiarization phase consists of logged descriptions of changes of scenes. The order of the scenes was the same for all subjects. The raw descriptions have been summarized into single labels that best describe the reports. The occurrences of these labels have been summed up per scene and over all scenes and are displayed in Table 4.1. The raw response logs can be found in Appendix A. Each column has to be interpreted as a report of change relative to the previous scene. Additionally, it is counted how many subjects detected a change of scenes in general and how many reports describe the intended change in perception (lower two rows). Except for the summed column, all numbers are limited to 11, i.e. the number of subjects. The underlined attribute labels are those, that have also been selected for the attribute rating task. No limit was set to the scope of reports, a subject was allowed to report as many attributes as they wanted. It shall be noted that the label *Distributedness* was used, when a subject describes the positions of speakers relatively to each other or to the listener. The label *Addressedness* describes an impression of a speaker being turned away/towards the subject. In general, all labels describe the impressions in both dimensions, meaning for example *Listening Effort* labelling the impression of something being perceived more and less effortful.

It can be noted that five of six attributes from the selected list of the rating task have been de-

Table 4.1 – Summarized outcome of the familiarization phase. The rows contain the attributes which ideally labelled the descriptions of the subjects for each change of scenes. The attributes are sorted by their occurrences. The lower two rows indicate if a change in scenes has been detected and if the intended perceptual change has been described. The underlined labels mark those, that have been selected for the attribute rating task. The last column sums up the occurrences over all scene changes.

Previous scene / **Current scene**	G1 K1	K1 K2	K2 G2	G2 G3	G3 K3	K3 K4	K4 G4	G4 G5	G5 K5	**Sums**
Distance	3	0	1	2	4	10	9	10	9	48
Distributedness	2	11	2	5	4	2	3	2	3	34
Reverberance	5	2	5	0	2	2	4	2	4	26
Clarity	1	0	0	0	2	5	3	6	4	21
Loudness	2	0	3	3	2	2	3	1	0	16
Room Size	2	2	1	0	0	2	3	2	2	14
Listening Effort	0	1	0	1	1	0	2	2	2	9
Width	2	0	1	0	1	0	0	0	0	4
Addressedness	0	0	0	0	0	2	0	1	1	4
Depth	2	0	1	0	0	0	0	0	0	3
Naturalness	1	0	0	0	0	0	0	0	0	1
Change detected	10	11	10	8	10	11	11	11	11	93/99
Intention detected	7	11	6	5	3	10	5	1	5	53/99

scribed intuitively by the subjects. This confirms the validity of the attribute list to be descriptive for the scene variations. The only missing attribute is *Realism*, which provides a first hint that no impression in the scene is intuitively described that way. Interestingly, *Distance* is described quite often, despite being only changed between scenes K3 to K4. It seems that the room change is often described with a change in perceived distance. As the room changes every second scene, the attributes *Reverberance, Room Size* and *Distance* are named most frequently. *Clarity* and *Loudness* are also prominent in the list. These attributes are mostly descriptive to the underlying changes in the signals due to e.g distance or facing angle changes.
Regarding change detection, the subjects have mostly described a change over all scenes. This number though underlies some bias, as the subjects are aware of when a scene is changed and are allowed to listen into the current and last scene in a pairwise manner. This measure has been conducted in order to see if the intended changes are detected. Obviously, the change of facing angle (G4 to G5) is not described as a change of *Addressedness* but more frequently as a change in *Distance*. This shows a clear tendency, that subjects are not able to detect the facing angle change in this setup. Note, that the subjects did not know that changing the speaker's facing angle was one possible variable. After this scene change, six of eleven subjects described a change in *Clarity*, which provides a clue, that the subjects detected the underlying change in the signal correctly (being comparable to a lowpass-filter), only the interpretation is wrong.
In general, for changes between the distant scenes, the real changes did not seem to have been detected correctly. This might be interpreted as some insecurity concerning distant sound sources. Also, the split of source positions between scenes G2 to G3 has not clearly been detected. Three of eleven subjects did not perceive a change at all.

4.2 Attribute Rating

The attribute rating task was the most extensive part of each appointment, as it took 42 Minutes on average of 81 Minutes duration of each appointment on average. Consequently, the aim of the analysis is focussed on reduction of the attribute list and thus the scope of the experiment. The quality of an attribute in this experiment is determined by its effect size regarding the conditions, redundancy compared to other attributes, the reliability and the consensus among subjects.
In Figure 4.1, all test ratings are scattered versus all retest ratings. The red line indicates the linear regression function. It can be noted, that there are ceiling and floor effects in the ratings. This reflects the tendency of the expert listeners to make use of the whole scale. $10.8\,\%$ of all ratings are classified as ceiling or floor data, which led to the decision not to perform compensatory transformation. A Kolmogorow-Smirnov-Test on the raw ratings of test and retest rejected the null-hypothesis of the data coming from a normal distribution at the $95\,\%$ confidence interval. The Spearman-correlation coefficient of 0.701 indicates a fairly strong correlation and thus

a good test-retest reliability.

An overview of the ratings of each scene is shown for each attribute in the boxplots of Fig-

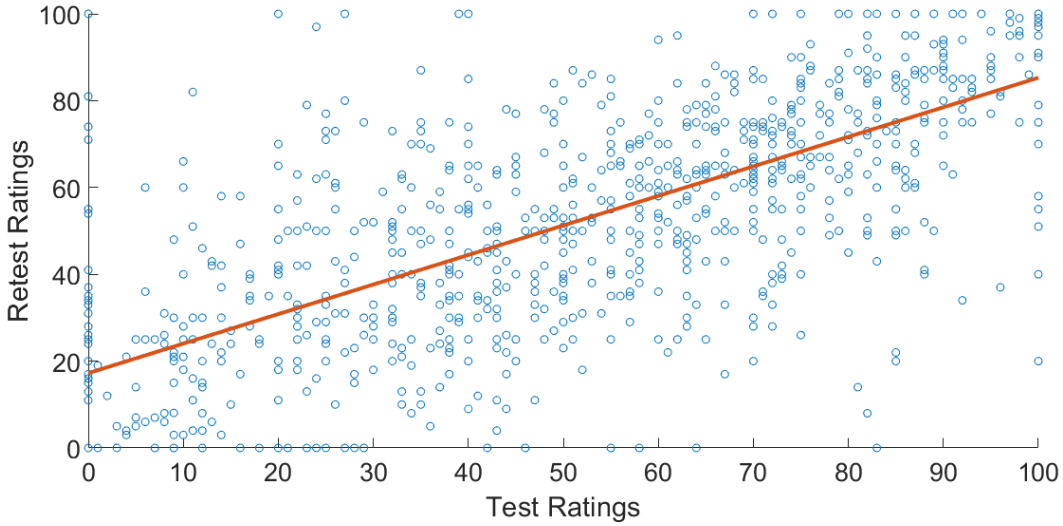


Figure 4.1 – Scatter plot of all test ratings versus corresponding retest ratings. The red line indicates the linear regression function, that was computed based on least squared errors to all data points.

ure 4.2. The blue boxes represent the test data, the red ones indicate retest data. Note, that each box consists of 11 ratings (one per subject). This display primarily serves as an overview to get a first impression of the shape of the data and to show trends that have to be verified in further analysis. It can be seen that the ratings of the attributes *Distance*, *Room Size* and *Distributedness* seem to spread less than that of the others. The ratings of *Realism* and *Listening Effort* are spread much more.

The trends in the ratings over the scenes show some resemblance for *Distance, Reverberance* and *Room Size*, indicating potential correlation, i.e. redundant information between them. The trends in the ratings of *Listening Effort* and *Realism* seem to be negatively correlated with respect to their variance, meaning that scenes with a high rating of realism are perceived as less effortful and vice versa.

The boxplots also allow to review some interpretations from the outcome of the familiarization task. It was assumed that the facing angle changes in scenes have mostly been responded with a perceived change of distance. This phenomenon can be further analysed looking at the attribute *Distance*. The assumption that facing angle changes are confused with distance seems to be confirmed looking at scenes #4 to #5 of the same room. In this part of the experiment the subjects were aware that the facing angle is one parameter that can change. The high variance in the *Realism* ratings can be considered a confirmation that it was not intuitively reported in phase 1. It is interesting to check the correspondence between real physical changes in the scenes and rated attributes. Technically, the ratings of *Room Size* and *Distance* were expected to cluster in two ratings, which is obviously not the case.

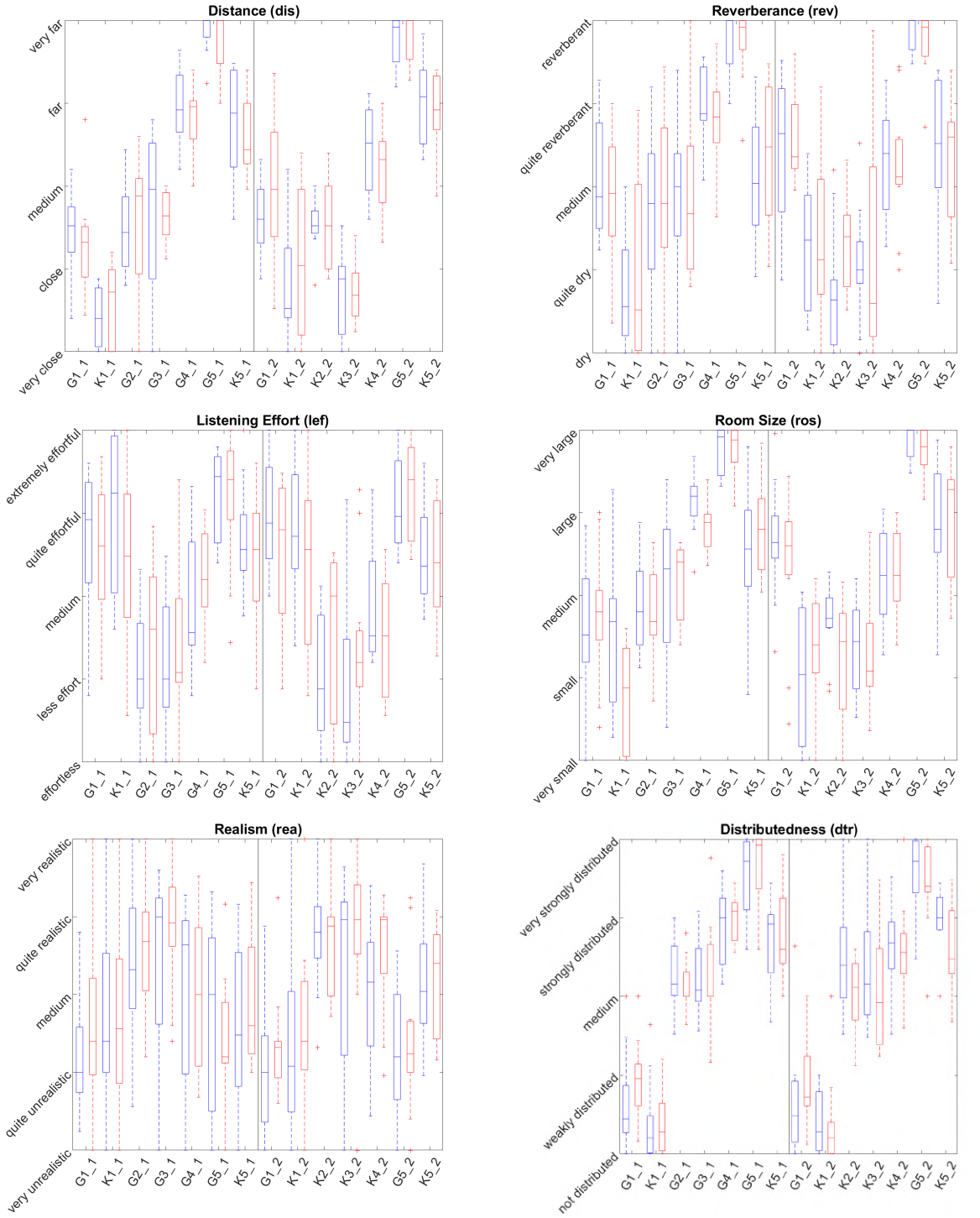


Figure 4.2 – Overview of the ratings over conditions for all six attributes. The blue boxes represent the test data, the red boxes represent the retest data. The scale labels are translations of the original German ones and are displayed on the values 0, 25, 50, 75 and 100. The markings on the abscissa represent the presented scenes. G and K indicate the rooms, the first index the scene and the second index the trial. The black vertical line separates both trials. The '+' symbols represent outliers.

To get a better overview of the data, a Principal Component Analysis (PCA) is computed in

Matlab [53]. Test and retest data have been treated separately. The result is displayed in the plots of Figure 4.3. In the right plot, each Principle Component (PC) is plotted corresponding to the variance that it explains. The blue line indicates the cumulative explained variance. In this analysis, only the first two PCs will be looked at, as they explain more than $75\,\%$ of the variance in the data.

In the left plot, the single data points indicate the variance of each sample group of 14 scene-ratings along the first two PCs. The data points are colour-coded for each attribute. The solid coloured lines point towards the centroids of the data points for each attribute. Additionally, the ratings of the ten scenes along the two PCs are displayed as grey for room Kircher and black for room Gravel. The scenes that have been rated in both trials (scenes #1 and #5) are represented by their centroid. It shall be noted that the PCA does not change the data, instead it only shows it from a different perspective. Most of the observations can also be made from the boxplots, while the PCA facilitates graphical inspection of the data.

The assumption from the boxplots, that the attributes *Distance, Reverberance* and *Room Size*

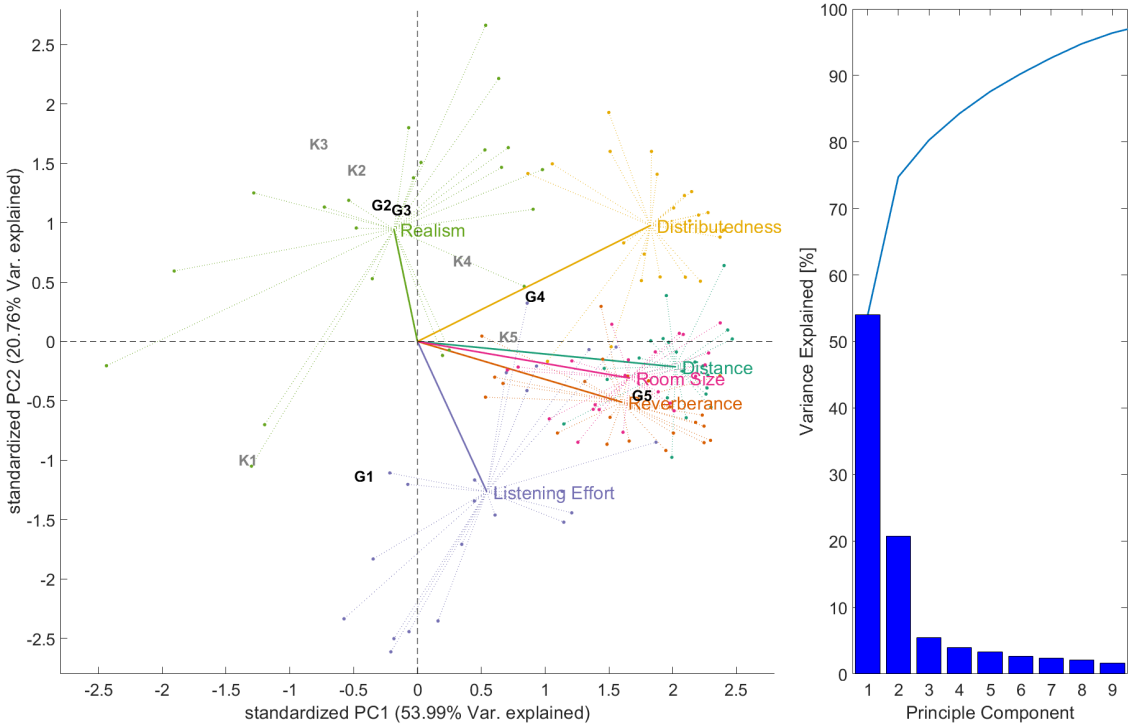


Figure 4.3 – Display of the result of the Principal Component Analysis (PCA). On the right, a bar plot is shown of the explained variance for each computed Principle Component (PC). The blue line indicates the cumulative sum of explained variance.
On the left, the data is plotted over the first two PCs. Each data point represents the variance in one subject's ratings along the two PCs. These data points are color-coded by their corresponding attributes. The vectors and their labels represent the centroid of each data cloud. In black and grey, the centroids of the scores of both rooms Gravel (**G#**) and Kircher (**K#**) are plotted.

are strongly correlated is supported by the small angles in between those attributes. They are also

close to the axis of the PC1, which indicates that this component reflects the room dimensions. Furthermore, the first PC explains 54 % in the ratings, which makes sense regarding parameters of room dimensions being most frequently changed throughout conditions. *Distributedness, Realism* and *Listening Effort* point in different directions, which is evident to the previous assumption of differing trends from the boxplots. *Realism* and *Listening Effort* seem negatively correlated among themselves. This also supports the observation of the trends in the boxplots, though again the high variance in these ratings have to be kept in mind. Kolmogorow-Smirnov-Tests on the raw ratings of each attribute reject the null-hypothesis of the data coming from normal distributions at the 95 % confidence interval. This taken into account, Spearman correlation coefficients between all attributes are provided in Table 4.2. Note, that these coefficients are calculated from the raw data, not the dimension-reduced PCA-scores. Thus, the interpretations can differ from the observations in Figure 4.3. The assumption of a strong negative correlation between *Listening Effort* and *Realism* from the PCA (large angle between the vectors) can be revised by a weak coefficient. The strong correlation between *Distance, Reverberance* and *Room Size* is confirmed. The PCA-scores of the conditions, i.e. the scenes give information about their ratings along the PCs. The scenes from room K are all positioned left and above from their equivalents in room G. Taking the attribute vectors into account, this reflects that room K has been rated with 'smaller' room dimensions (closer, less reverberant and smaller) and also as less effortful or more realistic. Looking at the scene indices along the first PC reflects greater room dimensions with greater indices, which is an expected outcome besides the confusion of distance and facing angle. Along the second PC it can be seen that scenes #1 are rated more effortful and less realistic. Scenes #2 and #3 are rated very similarly in both rooms which is evident to the assumed small perceptual differences in between.

The computation of the PCA can also serve the purpose to identify consensus on attribute rating among subjects. In Figure 4.4, the correlation of the raw ratings and their corresponding PCA-scores along the first two PCs is shown for each attribute. These types of plots are referred to as Tucker1- or correlation loadings plots [6]. The single data points in each plot indicate the correlation between the 14 raw ratings per attribute, subject and repetition and their corresponding scores of PC1 and PC2. In each plot, the same data points are displayed with only the ratings of

Table 4.2 – Spearman correlation coefficient matrix between the ratings of all attributes.
(* p-value < 0.05 ; ** p-value < 0.01 ; *** p-value < 0.001)

	Distance	Reverberance	Listening Effort	Room Size	Realism	Distributedness
Distance	1.000	0.650***	0.230***	0.758***	-0.115*	0.678***
Reverberance	–	1.000	0.295***	0.732***	-0.225***	0.466***
Listening Effort	–	–	1.000	0.285***	-0.363***	$-0.017^{(n.s.)}$
Room Size	–	–	–	1.000	-0.173***	0.571***
Realism	–	–	–	–	1.000	$0.100^{(n.s.)}$
Distributedness	–	–	–	–	–	1.000

each subject and their repeated ratings highlighted. The blue circles indicate how much variance in the ratings is be explained by the plotted components. The inner circle marks 50 % and the outer circle 100 % variance explained. Data points near the centre of the circles indicate, that most of the variance in these ratings is reflected by the other PCs.

The display allows interpretations about the inter-subject and intra-subject (test-retest) consensus. Data points that are close to another indicate good agreement. This can be seen for the attribute *Distance*, which clearly found most agreement among subjects. For *Reverberance* and *Room Size*, the data points seem to spread a bit more, while there are few subjects whose ratings could not be explained well by the two PCs. *Distributedness* is spread more, which indicates slightly differing comprehension of the attribute or its description among subjects. The variance in its ratings can be explained when the second PC is taken into account, which gives a another hint that *Distributedness* is rated based on different information than *Distance, Reverberance* and *Room Size*. *Listening Effort* clearly finds less agreement among subjects, although the variance in its ratings is mostly explained in the second PC. *Realism* shows least agreement among subjects and also among repetitions of the same subject, which is confirmatory to previous observations of this attribute.

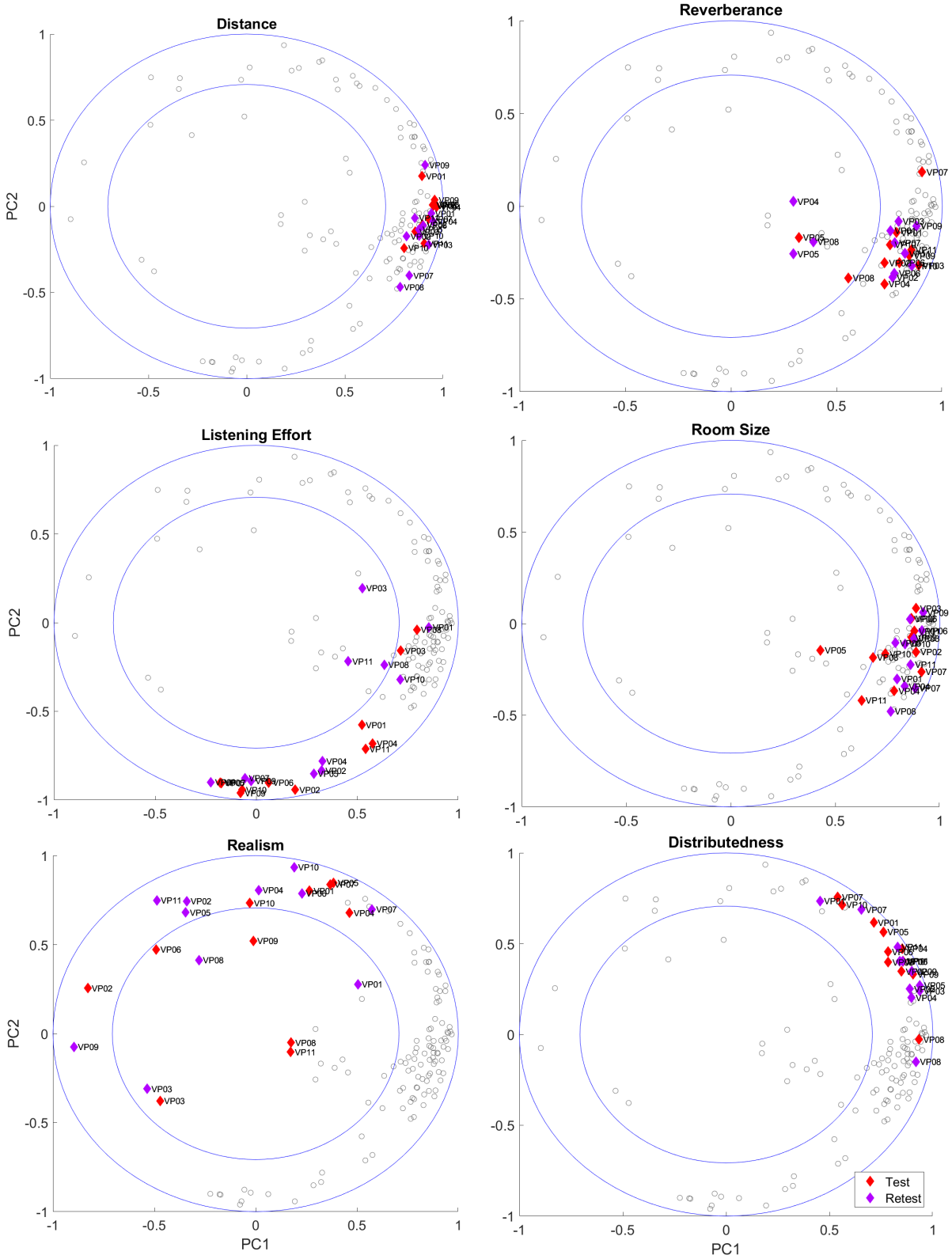


Figure 4.4 – Tucker1-plots of the Principal Component Analysis (PCA) for each attribute. The single data points in each plot represent the correlation between the raw ratings of one subject and the computed PCA-scores for the first two components. For each attribute, the data points of each subject and their repetitions are highlighted. The colours indicate test and retest. The outer blue circle indicates 100 % and the inner circle 50 % variance explained.

In order to attain quantitative information about the impact of single factors of the experiment, a four-factor ANOVA is computed in Matlab for each attribute separately. The four factors and their corresponding acronyms are the five Scenes (SCN), the two Rooms (ROM), the eleven Subjects (SUB) and the two Repetitions (REP). Effects of the scenes and rooms are the expected and thus fixed effects. Effects of the subjects and repetitions are considered unwanted and thus random effects. For easier interpretability, only two-way interactions between the factors are accounted for. A requirement of an ANOVA is that the residuals of the computed models have to be normally distributed. In Figure B.1 of Appendix B, the Cumulative Distribution Function (CDF)s of the residuals of each ANOVA-model are plotted versus the CDFs of the standard normal distribution. Additionally, a Kolmogorow-Smirnov test is calculated, which indicates normally distributed residuals in each case.
As a measure for effect size, $\tilde{\delta}$ are calculated for each factor, their two-way interactions and each attribute based on the following equation:

$$\tilde{\delta} = \sqrt{\frac{2}{n}}\sqrt{\frac{d.f.}{K-1}}\sqrt{F-1}$$

While *n* is the number of possible factorlevel combinations, *d.f.* are degrees of freedom of the factor, *K* is the number of levels in the factor and *F* represents the F-value. All values are provided in the ANOVA tables in Appendix B. According to «Sensory Evaluation of Sound» by N. Zacharov [6], $\tilde{\delta}$ can be interpreted as «*[...] the average pairwise difference between levels of a factor relative to the basic (residual) noise level in the data.*». $\tilde{\delta}$-values greater than 1 can be considered large effects, while it is advised to handle such global thresholds with care.
The fixed effects, meaning those of SCN, ROM or their interaction are displayed as bar plots in Figure 4.5. It can be seen that the scene variations had the largest effects on the ratings of all attributes. The interactions between the factors scene and room did not have noteworthy effects. The effect of the SCN on ratings of *Distributedness* are largest, which can be explained with it being the most frequently changed parameter throughout scene variations. Like in previous observations it is interesting that the ratings of *Distance* seem to be affected by scene variations despite the fact that only two levels of physical distance are applied. Looking at the effects of ROM, *Distance*, *Reverberance* and *Room Size* seem comparable, which indicates the relationship of distance perception and room characteristics. Interestingly, the effect of the room on *Reverberance* and *Room Size* is smaller than that of the scene. The effects of the scene on ratings of *Listening Effort* and *Realism* are considerably larger than those of the room. All interaction effects of scenes and rooms (SCN*ROM) are smaller than their single effects. Interactions of these factors could be interpreted as dependencies of the factors.

The random effects, i.e. the effects of repetitions, subjects, their interactions with each other and with the scenes and rooms are displayed in the bar plots of Figure 4.6. It can be noted that there are no significant effects of REP on the attribute ratings and no considerable effects of any

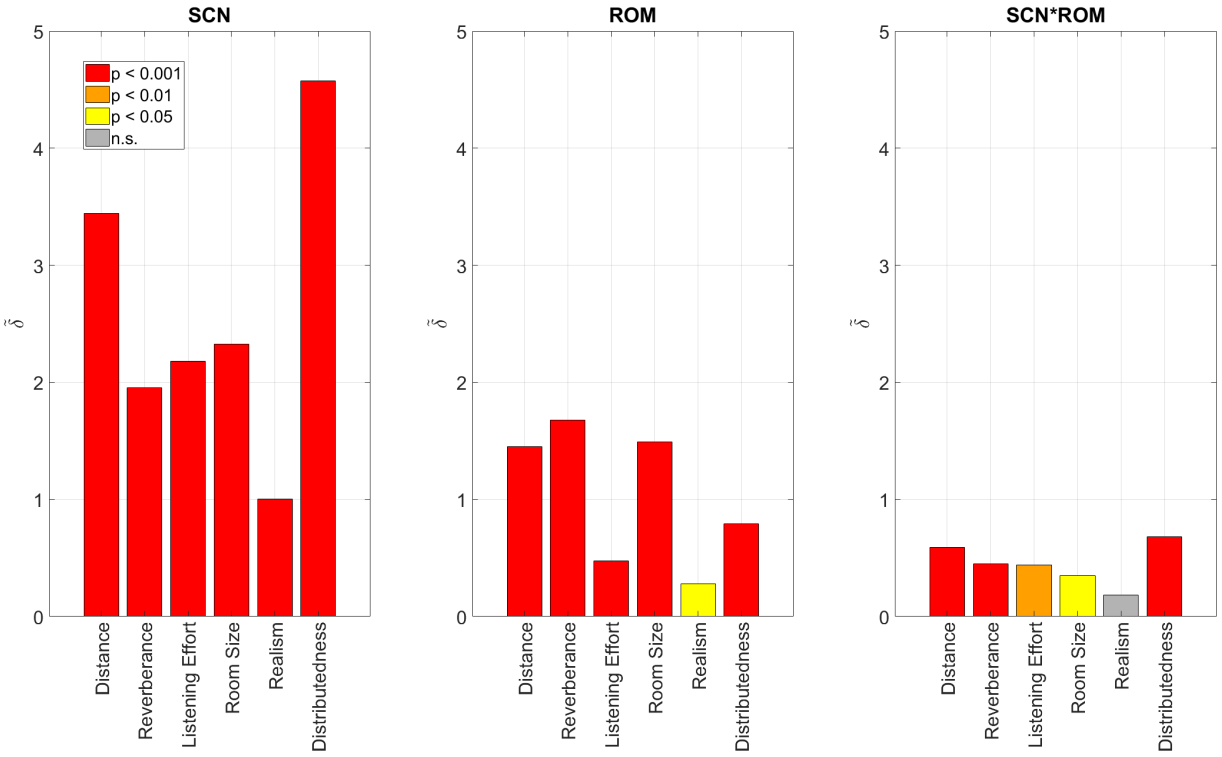


Figure 4.5 – Barplots of the effect size $\tilde{\delta}$ for each attribute on the factors Scenes (SCN), Rooms (ROM) and their interactions. Colour indicates significance of the effects.

interactions with this factor (SCN*REP // ROM*REP // SUB*REP). This supports the good test-retest-reliability that has been observed before. The effect of the SUB were largest for *Listening Effort*, which confirms the inconsistent consensus from the PCA. The interaction of subject and scenes (SCN*SUB) produced noteworthy effects on all attributes. This can be seen as the main source of noise in the data, i.e. the scene variations are interpreted differently between subjects, irrespectively of the attribute that is rated. The interaction of room and subject (ROM*SUB) shows slight effects for the attributes of room dimensions. The cause for that is seen in differing background of the expert listeners regarding room acoustics.

For more detailed information about the differences in the ratings of the single conditions and attributes, a post-hoc analysis is performed with the Matlab-function *multcompare*. The function is based on Tukey's HSD (honest significant difference) test. In Tables B.7 to B.12 of Appendix B, the results are displayed. The tables contain the differences in the means of the ratings in between all scenes of both rooms. Colour indicates significance. It can be seen from the tables, that the difference between scenes #2 and #3 is not reflected by any of the attributes in both rooms and all scenes (except for the scenes in room Kircher for *Distance*). The previous assumption of a facing angle change resulting in a change in distance rating can be confirmed, as scenes #5 have been rated significantly more distant than scenes #4. Scenes #4 on the other hand have been correctly rated as more distant than scenes #1 to #3.

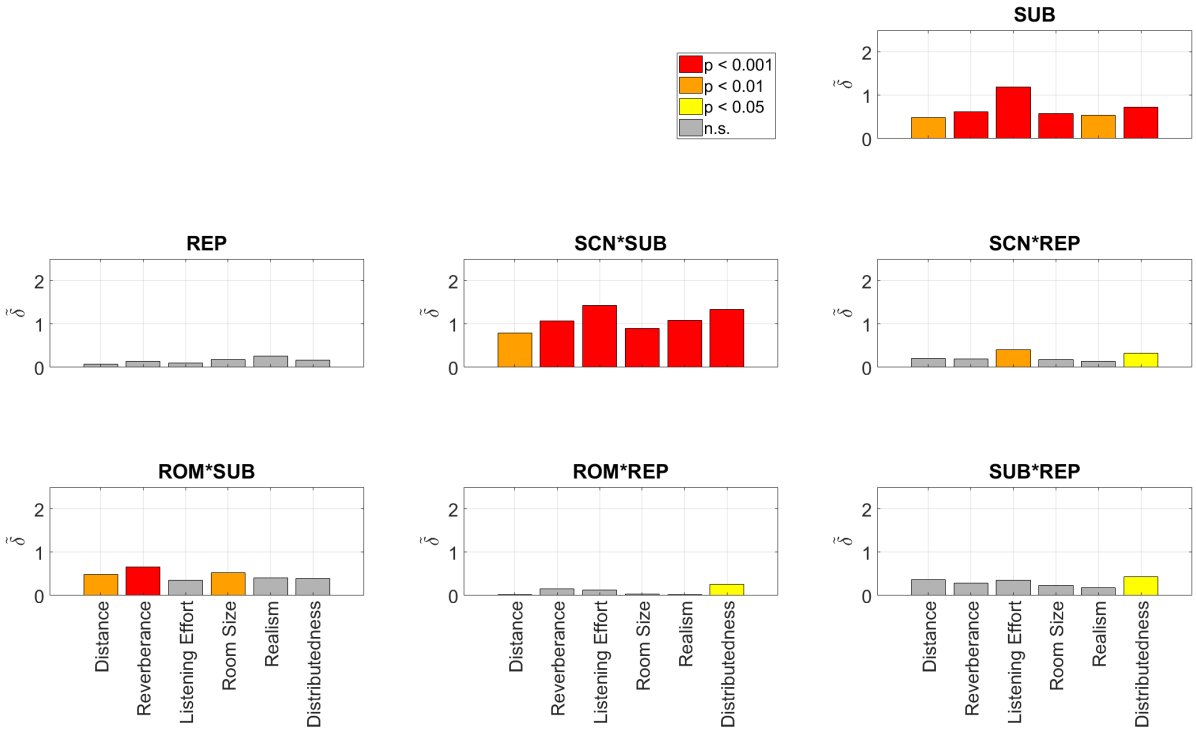


Figure 4.6 – Barplots of the effect size $\tilde{\delta}$ for each attribute on the factors Subjects (SUB), Repetitions (REP), their interactions and those with Rooms (ROM) and Scenes (SCN). Colour indicates significance of the effects.

From the bold values in the tables, the comparisons of the ratings of the same scenes between both rooms can be seen. The ratings of *Distance*, *Reverberance* and *Room Size* produced different ratings between the rooms for the most parts. *Listening Effort* and *Realism* does not seem to differ between the rooms. *Distributedness* only produced differences for scenes #1 and #5.

4.3 Figurine Alignment

The outcome of the figurine alignment task consists of 14 pictures per subject and per repetition. The pictures were made with a webcam from above the loudspeaker array. The first step for the analysis of this task is to digitize the position of each speaker for each alignment. This was done in a manual process by the investigator. The pictures were loaded into a Matlab interface which is shown as an example in Figure 4.7. On the left, a picture from the webcam is displayed along with a grid that serves as orientation and compensation for the camera misplacement. The picture is scaled from 0 to 1000, with 0 marking the origin (the room boundaries of the carpet grid) of the carpet grid and 1000 the outer limits of all figurine placements. In the right grid, the investigator clicks at the corresponding positions in the order of speakers 1 to 4 (top left to bottom right). The first click indicates the position, the second click places the vector that indicates facing angle. It has to be noted that this method of quantization adds noise to the raw data by the investigator

and can become quite extensive with larger data sets. The outcome of the digitization for each figurine consists of polar coordinates (lateral angle relative to the abscissa with its intercept at the listeners position in [°] and distance values in pixel units [p.u.]) and facing angles relative to the vector that connects figurine and listener. It shall be noted that no measure is taken to convert the arbitrary pixels into metric units as it is assumed not to reflect the relative distance ratios from the listeners. From the scaling process, the distance from the listener to the curtain (in the lab 1.2 m) equals 125 p.u.

An overview over the raw data is provided in Figure 4.8. Each plot contains the figurine align-

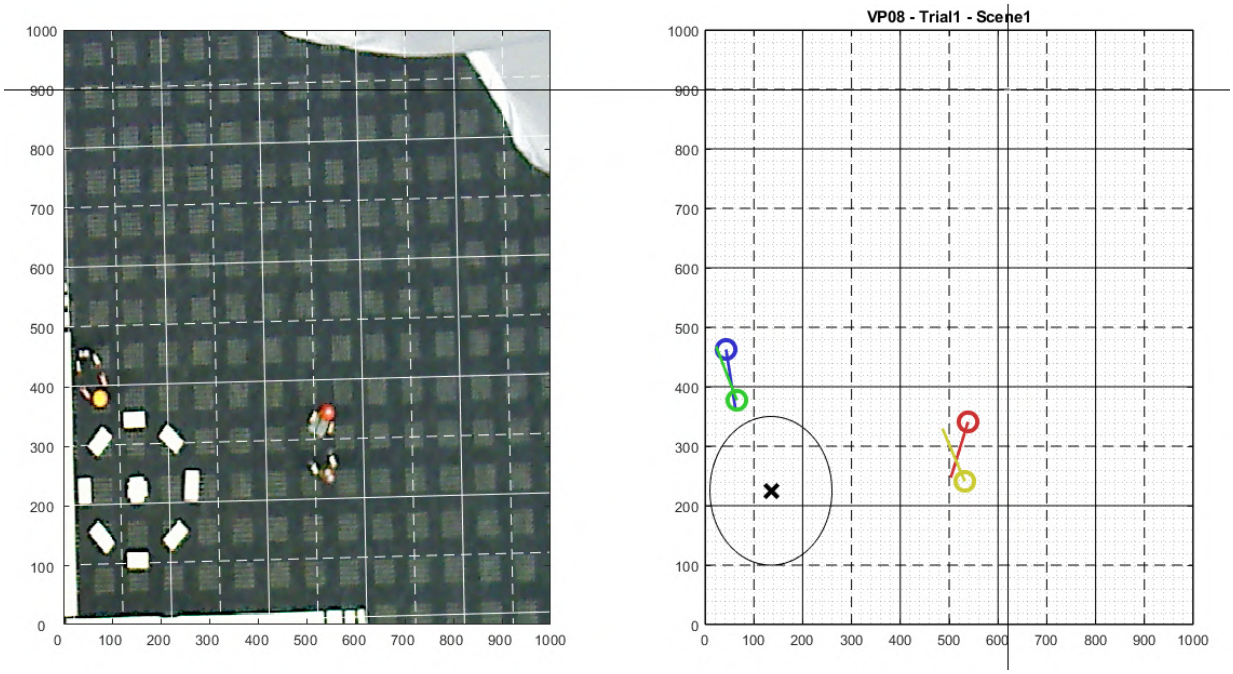


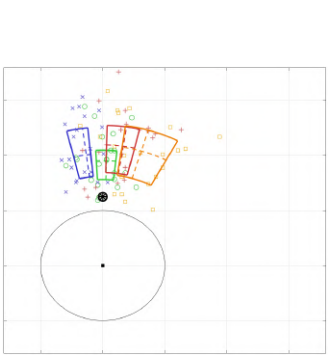
Figure 4.7 – Exemplary Matlab interface that is used to quantize the figurine positions. The left plot shows the picture from the webcam, along with a grid that equalizes the webcam misplacement and serves as orientation. The right plot shows the grid, in which the investigator clicked corresponding to the positions and facing angles of the speakers.

ment of one scene. Along with the raw positions, polar boxplots are displayed for each figurine. Each box represents upper and lower quartile of both lateral angle and distance. The lines within the boxes represent the median values. The black circle with the dot represents the position of the curtain and listener indicators from the carpet grid. The black symbols represent the ground truth position of each speaker. The ground truth positions are converted by the physical distances relative to the curtain distance from the speaker. This measure does not necessarily reflect the subjective distance ratio of the listeners, which is also why the distance measures are not converted into metric units. The grid in the plots is set to curtain distances to facilitate visual comparisons.

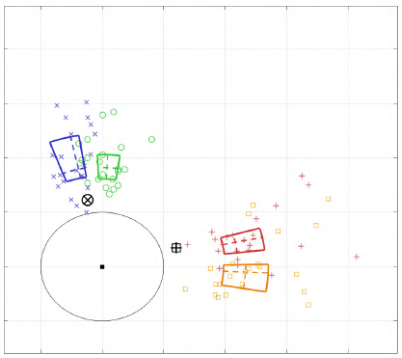
Like the boxplots of the attribute rating task, this display serves mainly as an overview of the data and to find trends that have to be verified in further analysis. A general observation is that

the positions of the scenes in room Gravel (upper rows) seem more spread than those in room Kircher (lower rows). It can also be seen that the positions of the right pair of speakers (red & orange) is spread more than those of the left ones (blue & green). Interestingly, the lateral angles of the right pair are not distributed around the ground truth but further right from it. It has to be kept in mind, that the subjects were generally situated such that the left speaker pair is in front. Although it was pointed out that they can and should turn their heads, the subjects were mainly facing the left pairs of speakers.

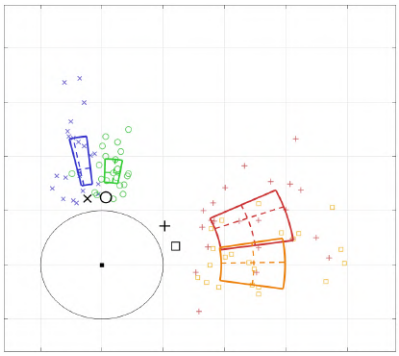
Comparing single scenes allows drawing assumptions about the perception of the physically conducted changes. The lateralization between scenes #1 and #2 seems correctly recreated. The lateral split of both pairs of speakers does not seem to have a big effect. It seems that in room Gravel, the split resulted in greater variance in the positions. The changes in distance between scenes #3 and #4 are also detected. In agreement to previous observations, the changes in facing angle are responded with changes in distance of the figurines and potentially larger variance in the positions in general.



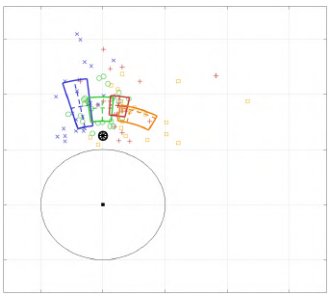
(a) Scene G1



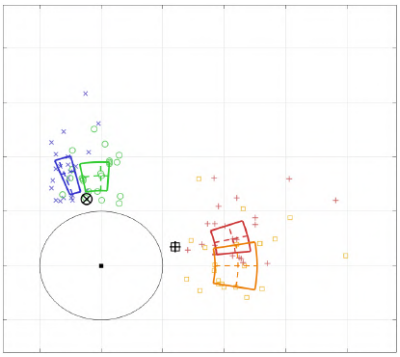
(b) Scene G2



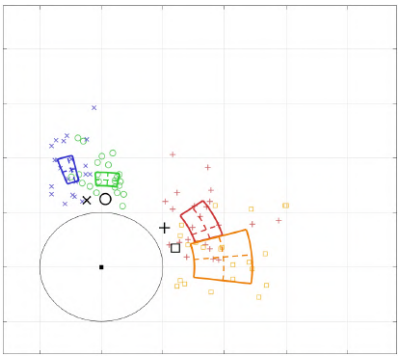
(c) Scene G3



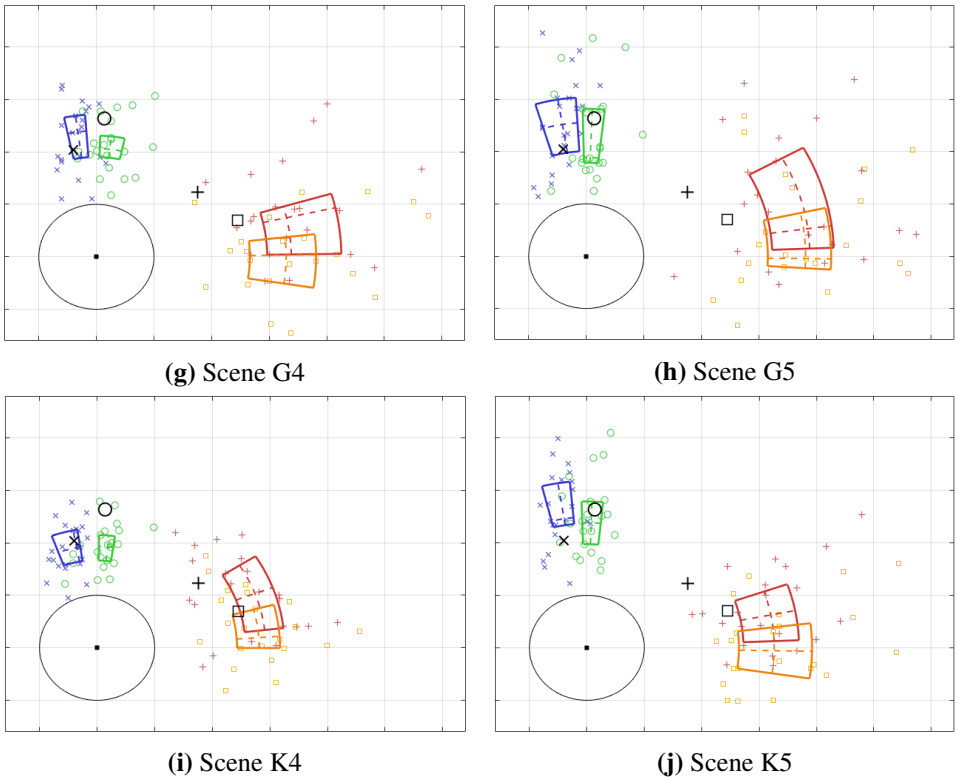
(d) Scene K1



(e) Scene K2



(f) Scene K3



(g) Scene G4 **(h)** Scene G5

(i) Scene K4 **(j)** Scene K5

Figure 4.8 – Polar boxplots for the figurine alignments of each speaker and each scene in both rooms. The boxes represent the upper and lower quartile for both lateral angle and distance. The dashed lines within the boxes represent their median. The single data points represent the raw positions. The colour-codes indicate speakers #1 (X), #2 (O), #3 (+) and #4 (□).

In order to attain qualitative information about the figurine alignment task, an ANOVA is performed. Unlike the four-factor-ANOVA from the attribute ratings, the outcome variables exist for each of the four speakers separately. Thus, the effects of the Speakers (SPK) are included as a fifth factor in the ANOVA. Instead of attributes, the outcome variables of distance (*Dist*), lateral angle (*Lang*) and facing angle (*Fang*) are analysed. In Figure C.1 of Appendix C, the residuals of the ANOVA models for all three variables are plotted as their CDFs versus the standard normal CDFs along with the results of the Kolmogorow-Smirnov Tests. The requirement for normally distributed residuals is fulfilled for *Lang* and *Dist*, while it is not for *Fang*. Consequently, the latter is excluded from the ANOVA and will be addressed separately. Before conducting the ANOVA, it has to be verified that the angular and distance errors are independent from one another. In Figure 4.9, the angular versus distance errors of all alignments are scattered. The weak Pearson-correlation coefficient of -0.2199 indicates independence of the two variables ($p < 0.001$), i.e. greater errors in *Dist* do not correspond with greater errors in *Lang* and vice versa.

In Figure 4.10, the previously introduced $\tilde{\delta}$ effect size measures are shown for the fixed effects of SPK, SCN and ROM and their two-way interactions. The largest effect on *Lang* is produced by SPK by far, which is consistent to the large angular distances in between scenes #2 to #5. Considerable effects are also observed for SCN and SCN:SPK. The largest effect on distance

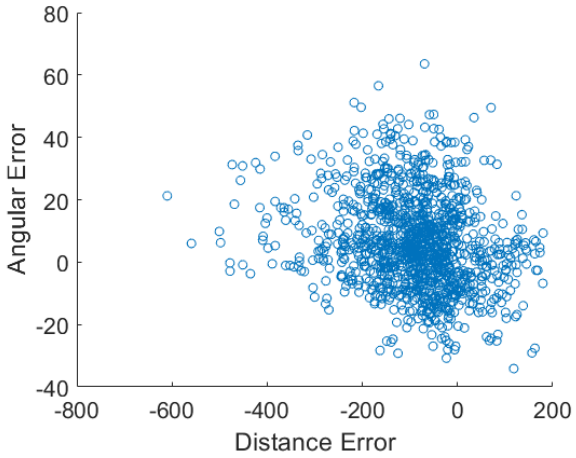


Figure 4.9 – Scatter plot of angular versus distance error from the figurine alignment task. The units of Distance Errors are pixel units [p.u.] from the quantization process. The Angular Error is displayed in [°].

placements are produced by SCN. In Figure 4.11, the $\tilde{\delta}$ values are plotted for the random effects SUB, REP and all remaining two-way interactions. No considerable effects come from REP. The most prominent observation from the random effects is that the SUB is a noteworthy cause of variance in the figurine alignment. This is interpreted either as a different conception of the speaker positions in the rooms or as an underlying noise due to differing size ratio concepts among subjects.

Like the results from the attribute rating task, Tukey HSD multiple comparisons tests are performed on the factor SCN and each Speaker separately. The results can be seen in Tables C.3 to C.6 of Appendix C. It can be observed that the *Lang* placements of Speakers 1 and 2 do not differ significantly in any of the scenes. These Speakers have only been moved in their distances in scene #5. Speaker 1 has been placed further away from the listener than Speaker 2, which is additional evidence to the facing angle and distance confusion. In the *Lang* placements of Speakers 3 and 4, significant differences compared to scene #1 can be noted, which is the most obvious change in between scenes. However, from the table it can be observed, that the mean *Lang* of scene #3 differs by 8.9° from that in scene #2. This shows a tendency, that the small difference between these scenes has been perceived at least for Speaker 3. From the differences in *Dist*, it can be seen that scenes #4 and #5 are correctly responded with larger distances in the figurines, though scene #5 is placed further away than scene #4. Interestingly, the lateral split of scene #2 is responded with greater distances for these speakers.

As a comparison between the figurine placements of the two rooms, the post-hoc tables with the factors SCN and ROM are displayed in Tables C.7 and C.8 of Appendix C. The bold values mark the comparisons of the same scenes between the rooms. A prominent observation is that for the most parts, the scenes in room Kircher have produced smaller distances in the figurine

alignments than those in room Gravel. The room-differences of *Lang* do not point out that much.

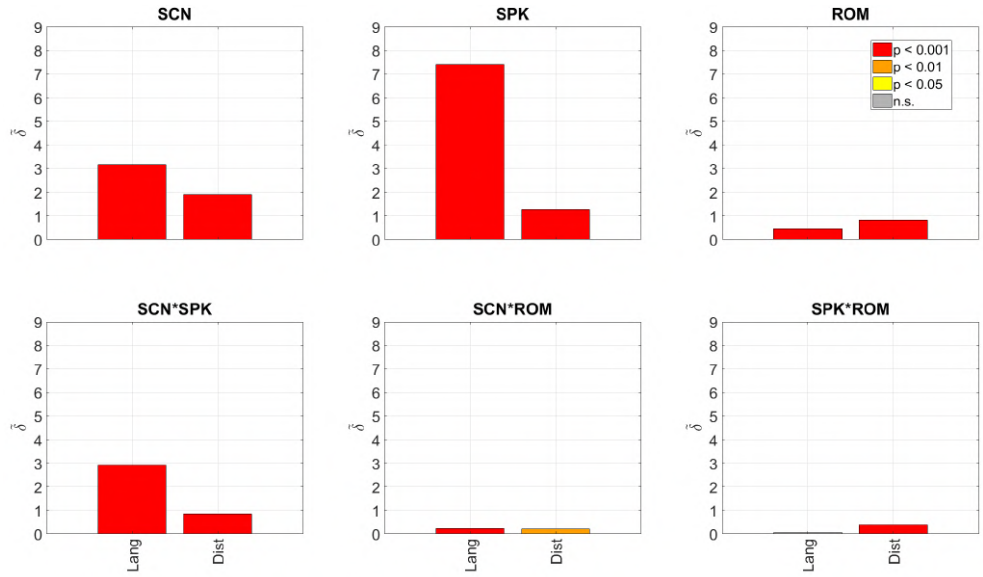


Figure 4.10 – Barplots of the effect size $\tilde{\delta}$ for errors in Distance (Dist) and Lateral Angle (Lang) on the factors Scenes (SCN), Rooms (ROM) and their interactions. Colour indicates significance of the effects.

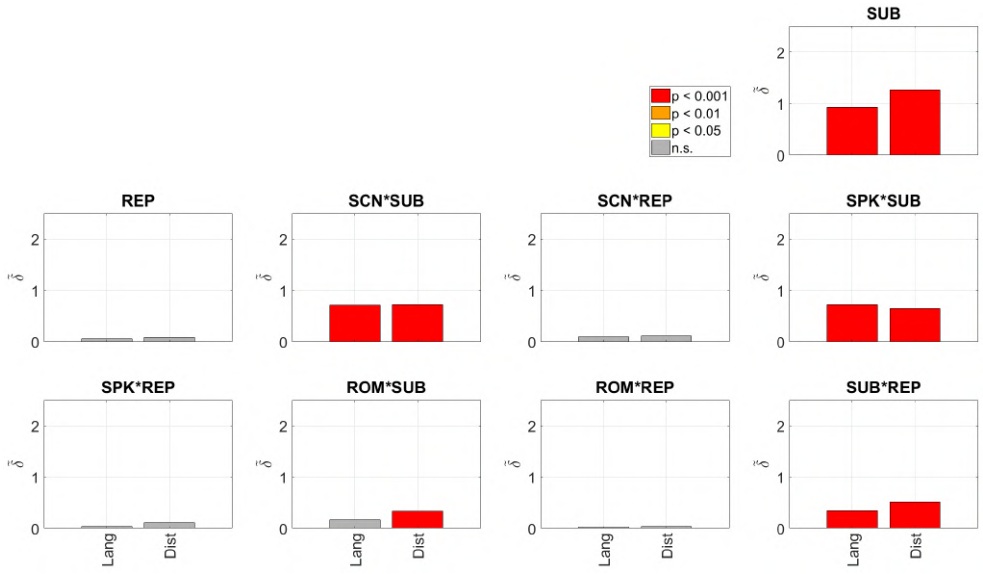


Figure 4.11 – Barplots of the effect size $\tilde{\delta}$ for each attribute on the factors Subjects (SUB), Repetitions (REP), their interactions and those with Rooms (ROM) and Scenes (SCN). Colour indicates significance of the effects.

In Figure 4.12, all test data is scattered versus all retest data for all three outcome variables. For *Dist* and *Lang* the error measures are taken. For *Fang*, the absolute data is displayed. The Pearson-correlation coefficients are $r_{Dist} = 0.66$, $r_{Lang} = 0.66$ and $r_{Fang} = 0.22$ with all

$p < 0.001$. The high correlation for *Dist* and *Lang* is evident to a good test-retest reliability, which can not be seen for *Fang*. The scattered data for *Fang* also reveals much noise in the data, which makes it unsuitable for any statistical analysis. A prominent observation from this plot is the higher density in data points towards zero.

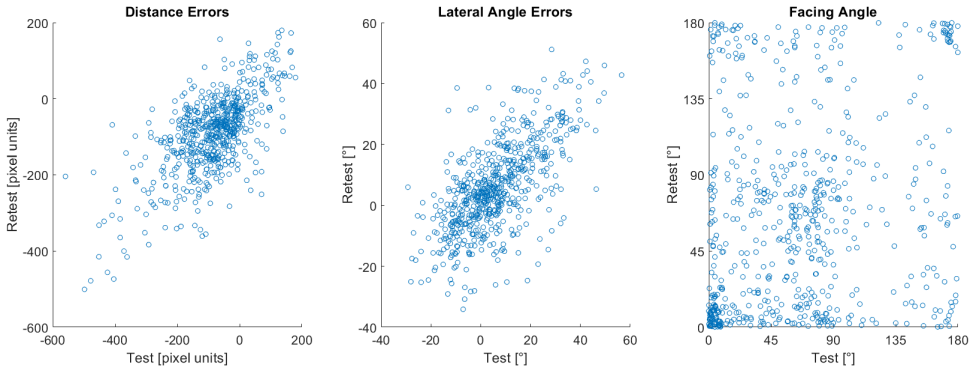


Figure 4.12 – Scatter plots of Test versus Retest data of the figurine alignment task. For *Dist* and *Lang*, the error measures are plotted. For *Fang*, the absolute facing angles are displayed.

In order to get a simplified overview of the data in facing angle placements, a simple classification is performed. Figure 4.13 shows a schematic of the classification. Three classes were defined as Facing Partner, Facing Listener and Facing Elsewhere. The partners position is also taken from the figurine placement. A threshold of $\alpha = 30\,^\circ$ is set as maximum deviation from Listener or Partner to determine the 'elsewhere' direction.

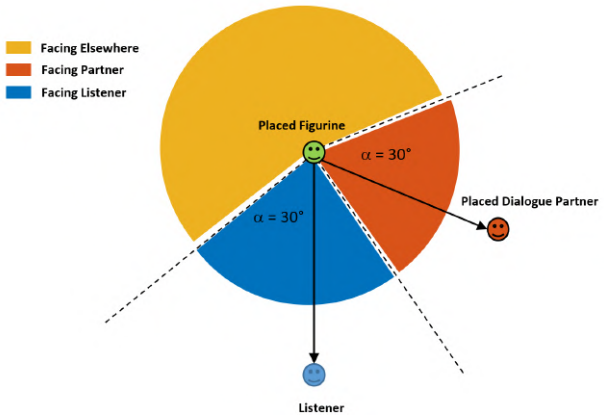


Figure 4.13 – Schematic of the classification of facing angle direction. The three classes were defined as Facing Listener, Facing Partner and Facing Elsewhere. α determines the maximum deviation from the partner or listener for the classes and is set to $30\,^\circ$.

In Figure 4.14, bar plots of the facing angle classes for each speaker are displayed. The bars show the proportion of facing angle classes relative to the amount of figurine placements. They

are split into the proportions of scenes #1 to #4 (where all speakers are facing the listener) and scenes #5 (where all speakers are facing their dialogue partners). Correct facing angle detection would be indicated by large blue bars in the scenes #1 to #4 (on the left) and large red bars for scenes #5 (on the right). Obviously, the ground truth data in facing angles is not reflected in the results. Only for speaker 3, the most prominent classes are the correct ones for the two conditions. For speaker 1 it can be observed, that the change in facing angle evoked an increase in the proportion of «facing elsewhere»-class. This is probably due to speaker 1 being the most turned away from the listener in scene #5, i.e. the effect of the low pass filter is largest. In general, the figurines have often been placed facing their partners, which can be considered a logical decision regarding the context.

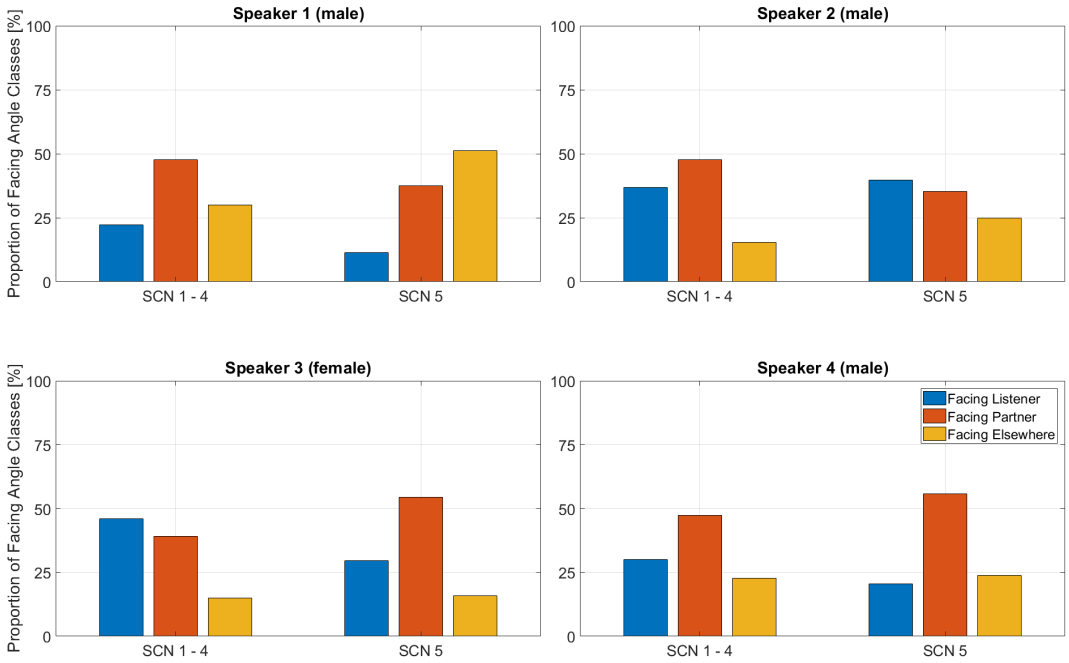


Figure 4.14 – Bar plots that represent the proportion of the three facing angle classes in the figurine alignment task for each speaker. In each plot, the left bars represent scenes #1 to #4, i.e. where all speakers truly faced the listener. The right bars represent scenes #5, i.e. where all speakers truly faced their conversing partner. Colour indicates the identified classes Facing Listener, Facing Partner and Facing Elsewhere.

4.4 Room Characterization

The outcome of the room characterization task consists of logged descriptions of scene #5 of both rooms for each subject. Additionally, the subjects had to assign each room one of four pictures that were layed out in front of them (see Figure 3.2). This task was conducted after all other tasks. The raw response logs are provided in Appendix D. The task was designed in order to attain information about the imagined rooms, which can not be seen from the figurine alignment task.

The logged descriptions of the two rooms can be seen as a confirmation to the hypothesis, that the room dimensions are overestimated compared to the ground truth. The rooms are described as bigger and more reverberant as they truly are. Room Gravel was frequently described as a large hall with plain surfaces and no obstacles inside. Room Kircher was frequently described as smaller with better acoustic design, i.e. less reverberant. The room was often described as a meeting room.
In Figure 4.15 the result from the forced choice task is shown as a bar plot. A prominent result from this task is that 10 of 11 subjects assigned room Gravel the picture of the big hall (Quattro). 7 of 11 subjects chose the picture of room Gravel as room Kircher. At least 3 of 11 subjects assigned the room Kircher the right picture. There is not enough data to this part of the experiment to draw global conclusions, though the result is confirmatory to the overestimation of room dimensions for room-in-room playback.

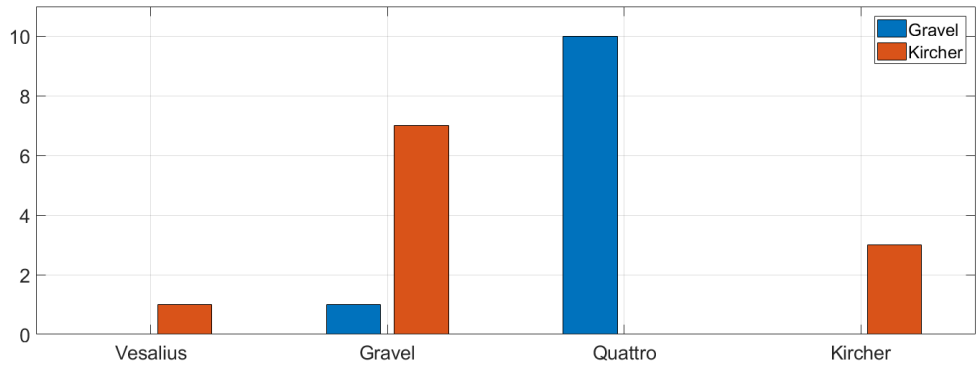


Figure 4.15 – Bar plot of the forced room choice task. The blue bars indicate the selections for the presented scene in room Gravel, the red bars indicate those for the scene in room Kircher. The Labels on the abscissa equals the names of the rooms where the pictures have been chosen from.

5 Discussion

The aim of this work was to find methods that are suitable to assess spatial sound quality with hearing aids. It is assumed that the ideal setting of hearing aid evokes a natural perception of space. Natural perception of space is considered to be equal to that of a normal hearing person. In this experiment, five scenes have been created in two rooms as experimental conditions. Ideally, all conditions are perceived the same among all subjects. Four tasks have been conducted that should ideally reflect the spatial images as specifically as possible. The outcome of the experiment yields much information about spatial perception in general, though the focus lies in rating the suitability of conducted tasks and experimental conditions.

Familiarization Phase The familiarization phase fulfilled its purpose to accustom the subjects to the presented materials and to sensitize them to the spatial aspects of the scene and room variations. The logged responses of the subjects to the changes contain much information about individual and unbiased comprehension of the scenes. They can be used to identify attributes that ideally describe the conducted scenes and their agreement among subjects.

Attribute Rating Task The suitability of the attribute rating task to measure spatial perception is dependant on the list of attributes that were rated. Here, it is of interest to identify attributes that do not add information or introduce noise to the data and revise the list accordingly. In this study, all subjects were trained and capable to conduct such an extensive rating task. The task took roughly two thirds of the whole experiment duration and was reported to be exhausting. Consequently, reducing the scope of the experiment is of interest especially when the subjects are not as trained as the expert listeners.
The results of the attribute rating task revealed that the ratings of the attributes *Distance, Reverberance* and *Room Size* are very similarly and well correlated. This indicates low information gain by letting subjects rate all three of them. From the PCA and the ANOVA it can be concluded, that *Distance* found the best consensus of all three and had the highest effect on the presented scenes.
Realism was neither rated reliably, nor found good consensus among subjects. The high overall variance in its ratings and the fact that there was almost no according description in the familiarization task provide reason for a rejection of the attribute. The reason for the poor result of *Realism* is believed to lie in the scene conception itself. Although the scene variations are designed as a thought of going from least to highest ecological validity, this does not imply realism. Physically, all scene variations can be considered realistic regarding their recording setup. Scene #1 for example is realistic if it is thought of one loudspeaker in a room that plays conversation

sound files. Perceptionally, all scene variations can be considered unrealistic, as the subjects are aware of the test chamber and its characteristics. Scene #4 for example is unrealistic regarding distances of single speakers that exceed the room limitations in the test chamber. Reviewing the literature on realism or plausibility as a percept, it is described as the agreement of the perception of the test material with an inner reference or the subjects expectations towards a corresponding real event [54],[40]. Such a reference is not provided in the instructions to this attribute, so the listener is forced to make one up. This probably led to the low agreement on the ratings of this attribute.

The attribute *Listening Effort* was rated differently from the other attributes with a tendencial negative correlation with *Realism*. The consensus among subjects was not ideal, it is the attribute with the highest effect from the subjects. Though the attribute seems to contain different information about the scene playback than the other ones and was rated reliably, which is reason to keep it in the list.

The ratings of the attribute *Distributedness* were also correlated with the room dimensions, though there is a trend that it contains additional information. The effect size from the scene variations was largest.

The recommended revised list of attributes for this experiment consists of *Distance, Listening Effort* and *Distributedness*. It has to be noted, that such attribute ratings have to be designed specifically depending on the products at test. Regarding hearing aid algorithms to be tested on their influence on spatial perception, the list might be expanded or reduced. For example, the implementation of hearing aid algorithms as test conditions along with the unprocessed scenes might provide an inner reference regarding the attribute *Realism*. This experiment also did not provide evidence, that the attributes *Distance, Reverberance* and *Room Size* really contain no differing variances in other contexts. From the familiarization task, it could be seen that attributes like *Clarity* or *Loudness* were frequently mentioned. Although they do not describe spatial aspects, they still provide information about the signals themselves, which can allow conclusions about the perception of spatial aspects to be drawn.

For a repetition of this experiment, it is also recommended to perform measures that avoid ceiling and floor effects to facilitate statistical analysis. To avoid such effects, the scale limits could be opened up or the scale labels could be positioned within the numerical limits. Alternatively, the scale labels can be restated so that their meaning exceeds the expected perceptual limits. For e.g. *Reverberance* the limiting scale labels could be revised to «very dry» and «very reverberant». Though it has to be kept in mind that this measure does not work if the subjects make use of the whole scale.

Generally, the attribute rating task was very extensive and effortful for all subjects. This is partly due to not setting up a reference scene, hence asking for all scenes to be rated relatively to all others. To facilitate this, the rating GUI could be expanded by a sorting function, which would

bring all the scenes that have already been rated in an ascending order. At the example of *Realism* it is recommended to refine attribute introductions if they are not as unambiguous as e.g. *Distance*. For that attribute, the refined instructions would need to set up an inner reference that the subjects agree on.

Figurine Alignment The figurine alignment task was the attempt to create an experiment with which the spatial aspects in the scenes can be assessed non-verbally. A great advantage of this task is that it does not rely on verbal agreement of proposed attributes, their labels or their definitions. It can be expected that the task can be implemented easily in different countries with a good comparability, which is more difficult with verbal elicitation techniques [42]. The task also has potential in the assessment of the perception of space in children as it can be set up in a playful way. Another advantage comparing to the attribute rating task is that it provides information about the perception of the single speakers in the conducted scenes. Especially when it comes to more complex presentations of spatial information, it is hard to make sure that the attributes are rated based on all presented elements. In this experiment, the listeners were instructed to take all speakers into account for the attribute ratings, though there is no way to control this in the outcome. In the figurine alignment task, the subject is forced to analyse each speaker separately. It could be seen from the ANOVA, that largest effects on the distance and lateral angle placement can be attributed to the single speakers. Such effects can be assumed to be present in the attribute rating task as well, except that there is no way to prove this. The figurine alignment task took 15 Minutes on average, while the attribute rating task took 42 Minutes on average.

Another considerable effect on the variance in distance and lateral angle placement was produced by the subjects. It is assumed that this effect is mainly caused by the differing perceptions of size ratios. The distances of the markings on the carpet were made based on downscaling the relative size of a figurine to the curtain distance. It is conceivable, that each subject has a different perception of such relative distances. This can be avoided by implementing training sessions for this task to equalize the subjective distance ratios among subjects. An interesting observation from this task was that figurines 3 and 4 were placed more distant after the lateral angle split. The cause for this is assumed to lie in a differing distance perception due to the absence of the closest walls in comparison to scene #1. It is also conceivable, that the subjects did perceive the same distance as in scene #1, just the placement underlies distortions due to a transformation from a self-centred acoustic perception to the isometric perspective on the carpet grid.

The grid on the carpet marks one general flaw in the design of this task. It requires the subjects to bend out of the HOA-sweet spot, thus distorting the auditory image of the scenes. Furthermore, bending over to reach the floor requires a certain level of manoeuvrability, which cannot be de-

manded from older subjects. It also requires the ability from the subjects to be able to abstract all auditory room dimensions to a visual domain in a perspective from above. It is unclear, if this perspective is also distorted when further abstracting it to the grid on the carpet. In the polar boxplots it was observed, that the lateral positions of the right pair of speakers was not distributed around the ground truth but even further right from it. As there are many sources of distortion in the image, it is not clear what causes this effect. The generally larger spread of data for the right pair of speakers can be explained by decreasing accuracy in azimuth localization towards the sides [55].

An alternative approach to the carpet grid would have been to insert a small table next to the subject with the same grid. One problem of this approach is, that any additional obstacles such as a table in the room can interfere with the HOA-playback, thus distorting the spatial impression within the sweet spot. Another problem would have been that the table area is limited, which also limits the dimensions of the figurine placement and implies a certain room size. The subjects should place the figurines based on reference distances of what they see, meaning the curtain distance as primary reference. If they perceived the speakers very far away, a table would prevent them to place the figurines accordingly.

Another alternative could be seen in a study by Hassager et al. [56]. In that study, the subjects were instructed to position sound sources on a digital grid on a tablet in front. The problem of a reflective obstacle remains, though it might have less of an impact compared to a table. The scales on a digital grid could be changed by the subject, which could solve the problem of boundaries. A new issue with this method might be that pointing dots on a digitial grid requires a better ability of spatial abstraction in the subject, while handling figurines is learned already as a child.

A solution for the distortion in perspective, which would not interfere with the HOA setup could be the application of virtual reality glasses. Although this approach would be costly, it would preserve the self-centred perspective. Subjects would position human-sized avatars into a virtual environment, until the visual input matches the auditory. Additionally, a virtual environment might lead to a better perceptual deprivation of the real measurement chamber. Apart from the expenditure to implement such a method, it is considered not to be suitable for older subjects.

The figurine alignment task allowed very specific conclusions about facing angle perception to be drawn in this experiment. From the other tasks only hints have been observed that facing angle changes are not perceived as such. In the instructions of this task it was pointed out that the facing angle of the figurines will be analysed as well. In the attribute rating task it was revealed that the facing angle is one possible parameter that changes in between the scenes. Despite this, no clear results could have been drawn regarding facing angle placements. The test-retest-scatter plot (Figure 4.12) indicated rather random placement, the bar plots (Figure 4.14) only allowed the assumption that the single speakers produced different facing angle placements, though this is again not related with their real facing angle. From the familiarization task it could be seen that

at least the effect of the spectral tilt has been detected, as a change in clarity has been reported. While it has been shown that facing angles of competing talkers affect the speech perception in complex scenarios [24], it can be hypothesized from this task, that blind facing angle detection in this form is not a needed ability in everyday life. Facing angle of sound sources has shown to be an ambiguous cue, which leads to the spatial context of the alignment to be pivotal on the facing angle perception.

Room Characterization The room characterization task allowed conclusions to be drawn about the perception of the rooms where the scenes are situated in. From both the detailed descriptions and the forced choice task, it could be clearly seen that the room dimensions of the scenes are overestimated. The rooms were described as larger and more reverberant than they truly are. According to a study by Yadav et al. [57], the percept of reverberance is an indicator for perceived room size. It is assumed that the estimation of the room dimension apart from the visual domain is triggered by the initial conversation between subject and investigator. In this study, the instructions to the experiment and additional conversation took place in the measurement chamber. It can be assumed that the perceptual room characteristics are set up already when entering the room and in conversations during instructions. The perceptual room characteristics that are suggested by the scene playback are deceiving the already defined room perception. In a study by Schutte et al. [58], visual room impression did not influence the subjective ratings of reverberance. The effect of overestimating the room dimensions can also be experienced when listening to the convoluted dialogues via headphones. After the experiment, the true rooms were revealed to the subjects. Most subjects informally reported that they would have expected less reverberant room characteristics from the real rooms.
Some subjects described their strategy during this task was to think of a real corresponding room that they know and describe the image of that room. When further asked about that strategy, it can be noted that these detailed imaginations of the rooms were not present before they were asked for it. It is assumed that these images of the rooms constitute a certain bias. It was intended to keep the subjects naive to the auditory nature of the scenes for as long as possible. It is conceivable, that such a room characterization task could influence the ratings for attributes of room dimensions if executed before.

Scene Validation It is of interest to reduce the scope of the experiment in order for it to be suitable for non-expert or older hearing impaired listeners. The scope can be reduced by applying the recommended revised attribute list as it was described before. Reducing the list of experimental conditions, i.e. the scenes becomes especially important when additional condi-

tions should be introduced. Such conditions could be certain hearing aid algorithms. From all of the four tasks, it is consistently observed that the scenes #5 were not perceived the way they were intended to be. This constitutes interesting new research questions regarding facing angle detection, though these scenes are assumed not to be suitable for the rating of spatial sound quality. In a repetition of the experiment it is advised to remove scene #5 from the list, although this is not consistent with the demand on the scene variations to be ecologically valid. Comparing the scenes to situations in real life, scene #5 can often occur. It is highly unlikely, that surrounding speakers are facing the listener, without addressing her/him in the conversations. In this context, scenes #5 might be perceived as more realistic when a fifth speaker is introduced that constitutes a desired target speaker for the listener. However, in a supposedly ecologically valid experiment, a target speaker that does not involve the listener in an active conversation is considered unrealistic. The tasks in this study were conducted, so that the subject is an uninvolved observer of the scene. In a comparable scenario in everyday lives, the conversing speakers are expected to be identified as irrelevant auditory objects, thus their acoustic details are not of greater interest for the listener.

Another observation was that the changes between scenes #2 and #3, i.e. the lateral split of the conversing partners were not clearly reported. This can be seen by the small distances of these scenes in the PCA-plot of the attribute rating task (Figure 4.3), from the reports in the familiarization task and the small changes in the figurine alignments. A revised list in for study consists of scenes #1, #3 and #4.

It has to be noted, that there are numerous speaker alignments to be thought of in such experiments. This study showed that the conducted scenes enable differentiable ratings of spatial attributes. To gain ecological validity, moving sound sources or ambient background signals have to be considered. The naturalness of the scenes is expected to be dependant on the quality of the dialogue sound files. The suitability of such sound files is dependant on many factors, such as their intonation regarding the surrounding scenario (Lombard Effect) and also the contents of the dialogue. The introduction of a target speaker that provides desirable information for the listener might increase the necessity of the listener to analyse spatial details of their surroundings. Furthermore, it would be interesting to analyse the effect of the amount of speakers in the scene on the accuracy in figurine placements or the attribute ratings. All in all, it should not be disregarded, that outside of such laboratory settings, the most important and reliable information about the surroundings comes from the visual domain. It is highly desirable to design experiments that include the visual domain while not constituting too much bias on the auditory perception.

6 Conclusion

In the here conducted experiment, several new methods were developed and evaluated. The overall motivation was to create an experiment that is suitable to assess spatial aspects of sound quality with hearing aid users. The conducted scenes were created using HOA reproduction. The technique itself is suitable to evaluate hearing aid features, although experiments with beamformers have to be handled with care. The HOA playback allows it to create perceptually plausible scenes, while still enabling natural sensory-motor feedback to be used as spatial cue. The scenes were created with impulse responses for each used position in the scenes. This method has the advantage, that these impulse responses can be convolved with different dialogues, with music instruments or ambient sound sources. Additionally they can be transferred to other HOA reproduction systems.

The design of the scenes was conceptualized to present spatial information in different degrees of ecological validity. It could be seen, that the most ecologically valid scene was not identified as such. This constitutes new research questions regarding facing angle detection and naturalness of such virtual acoustic environments. For the purpose of evoking different spatial perceptions, the scenes are considered suitable. Perceptual differences in the locations of the speakers and the single attributes have been observed for the two rooms.

The familiarization task contained a lot of information about the unbiased or naive perception of the presented material. From this experiment, it is recommended to conduct such a task, as the logged responses contain much information about the perception of the individual and unbiased subject. Though it would be interesting to see how the order of scene playback influences the perception of the differences.

The attribute rating task produced significant effects and provided much information about scene and room variations and the relationship of the single characteristics. However, the task was very extensive, which is motivation to reduce its scope. The attribute *Realism*, *Reverberance* and *Room Size* are recommended to be excluded from the list due to either unreliability or redundancies towards other attributes. The instructions to the attribute *Listening Effort* has to be revised, in order to gain consensus about the ratings. Training sessions to set a clear inner definition about the attributes are conceivable. Because of the ceiling and floor effects in the ratings, it is advised to open up the scales, allowing the ratings to exceed 0 and 100. Additionally, the scale end labels should be positioned further towards the centre of the scale. To facilitate this task, a reference condition could be implemented, so that non-expert assessors can rely on pairwise-comparisons.

The figurine alignment task is considered an explorative method to assess spatial quality non-verbally. The task itself can be considered to be intuitive for most subjects. It took roughly only one third of the time of the attribute rating task and contained additional information about the single speakers and their orientation. It is concluded, that the task has great potential in spatial sound quality assessment, though there is room for improvements. It should be changed, so

that subjects are not required to bend out of the sweet spot and over a comfortable level. The problem could be resolved with a GUI on a tablet or the figurines on a small table. Virtual or augmented reality glasses should also be considered. Additionally, to facilitate the analysis of the scene recreations, automatic ways to digitize the positions should be implemented. Furthermore, training sessions should be considered in order to equalize the differing distance and size ratios among subjects.

The room characterization task clearly revealed over-estimation of the room dimension. It motivates further research to find out the cause for this phenomenon. For application in hearing aid assessment, it is conceivable to present each room once with and without the active feature. It is recommended to conduct this task at the end of the experiment as the imagined rooms seemed to further establish during the descriptions. An established image of the virtual room is hypothesized to impose a certain bias on the room perception, although this also has to be further investigated.

References

[1] J. Blauert, *Spatial Hearing - The Psychophysics of Human Sound Localization*. The MIT Press, 2001.

[2] D.L. Blackwell, J.W. Lucas and T.C. Clarke, "Summary health statistics for u.s. adults: National health interview survey," *Vital Health Stat*, vol. 10, no. 260, pp. 1–161, 2014.

[3] T. Van den Bogaert, T.J. Klasen, M. Moonen, L. Van Deun and J. Wouters, "Horizontal localization with bilateral hearing aids: Without is better than with," *J. Acoust. Soc. Am.*, vol. 119(1), pp. 515–526, 2006.

[4] H. G. Hassager, A. Wiinberg and T. Dau, "Effects of hearing-aid dynamic range compression on spatial perception in a reverberant environment," *J. Acoust. Soc. Am.*, vol. 141, no. 4, pp. 2556–2570, 2017.

[5] I. M. Wiggins and B. U. Seeber, "Effects of dynamic-range compression on the spatial attributes of sounds in normal-hearing listeners," *Ear & Hearing*, vol. 33, pp. 399–410, 2012.

[6] N. Zarachov, *Sensory Evaluation of Sound*. CRC Press, 2019.

[7] S. Kochkin, "Marketrak viii: Consumer satisfaction with hearing aids is slowly increasing," *The Hearing Journal*, vol. 63, no. 1, pp. 19–32, 2010.

[8] M. Akeroyd, "Effect of hearing aids on distance perception," *J. Acoust. Soc. Am.*, vol. 123, no. 5, 2008.

[9] L. S. R. Simon, H. Wuethrich and N. Dillier, "Comparison of higher-order ambisonics, vector- and distance-based amplitude panning using a hearing device beamformer," *4th International Conference on Spatial Audio, Graz, Austria*, 2017.

[10] L. A. Jeffress, "A place theory of sound localization," *J Comp Physiol Psychol*, vol. 41, no. 1, pp. 35–39, 1948.

[11] M. P. B. Grothe and D. McAlpine, "Mechanisms of sound localization in mammals," *Physiol Rev*, vol. 90, pp. 983–1012, 2010.

[12] E. A. Macpherson and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *J. Acoust. Soc. Am.*, vol. 111, no. 5, pp. 2219–2236, 2002.

[13] R.J. Otte, M. J. H. Agterberg, M.M. Van Wanrooij, A. D. F. M. Snik and A. J. Van Opstal, "Age-related hearing loss and ear morphology affect vertical but not horizontal sound-localization performance," *J. Assoc. Res. Otol.*, vol. 14, pp. 261–273, 2013.

[14] P. Zahorik, "Auditory display of sound source distance," *Proceedings of the 2002 International Conference on Auditory Display, Kyoto, Japan*, 2002.

[15] S. Carlile and J. Leung, "The perception of auditory motion," *Trends in Hearing*, vol. 20, pp. 1–19, 2016.

[16] J. Lewald and H.-O. Karnath, "Vestibular influence on human auditory space perception," *J. Neurophysiol.*, vol. 84, no. 2, pp. 1107–1111, 2000.

[17] S. Carlile, "The plastic ear and perceptual relearning in auditory spatial perception," *Frontiers in Neuroscience*, vol. 8, no. 237, pp. 1–13, 2014.

[18] T. M. Woods and G. H. Recanzone, "Visually induced plasticity of auditory spatial perception in macaques," *Current Biology*, vol. 14, pp. 1559–1564, 2004.

[19] D. L. Valente and J. Braasch, "Subjective scaling of spatial room acoustic parameters influenced by visual environmental cues," *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 1952–1964, 2010.

[20] F. Dollack, C. Imbery, S. van de Par and J. Bitzer, "Einfluss von visueller stimulation auf distanzwahrnehmung und externalisierung," *42. Jahrestagung für Akustik DAGA, Aachen, Germany*, 2016.

[21] D. Cabrera, D. Jeong, H. J. Kwak and J.-Y. Kim, "Auditory room size perception for modeled and measured rooms," *The 2005 Congress and Exposition on Noise Control Engineering, Rio de Janeiro, Brazil*, 2005.

[22] S. Hameed, J. Pakarinen, K. Valde and V. Pulkki, "Psychoacoustic cues in room size perception," *AES 116th Convention, Berlin, Germany*, 2004.

[23] H. Kato, H. Takemoto, R. Nishimura and P. Mokhtari, "Spatial acoustic cues for the auditory perception of speaker's facing direction," *20th International Congress on Acoustics ICA, Sydney, Australia*, 2010.

[24] O. Strelcyk, S. Pentony, S. Kalluri and B. Edwards, "Effects of interferer facing orientation on speech perception by normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 135, no. 3, 2014.

[25] G. Grimm, B. Kollmeier and V. Hohmann, "Spatial acoustic scenarios in multichannel loudspeaker systems for hearing aid evaluation," *J. Am. Acad. Audiol.*, vol. 00, no. 0, pp. 1–10, 2016.

[26] C. Oreinos, *Virtual Acoustic Environments for the Evaluation of Hearing Devices*. PhD thesis, Macquarie University, Sydney, 2015.

[27] V. Pulkki, *Spatial Sound Generation and Perception by Amplitude Panning Techniques*. PhD thesis, Helsinki University of Technology, 2001.

[28] A. J. Berkhout, D. de Vries and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Eng. Soc.*, vol. 93, pp. 2764–2778, 1993.

[29] S. Moreau, J. Daniel and S. Bertet, "*d sound field recording with higher order ambisonics - objective measurements and validation of a 4th order spherical microphone," *AES 120th Convention, Paris, France*, 2006.

[30] G. Grimm, S. Ewert and V. Hohmann, "Evaluation of spatial audio reproduction schemes for application in hearing aid research," *Acta Acustica united with Acustica*, 2014.

[31] G. Grimm, J. Luberadzka, T. Herzke and V. Hohmann, "Toolbox for acoustic scene creation and rendering (tascar): Render methods and research applications," *Linux Audio Conference, Mainz, Germany*, 2015.

[32] Max/MSP, "https://cycling74.com/," Aug. 2019.

[33] Ircam Spat, "http://forumnet.ircam.fr/product/spat-en/," Aug. 2019.

[34] C. Oreinos and J. Buchholz, "Evaluation of loudspeaker-based virtual sound environments for testing directional hearing aids," *J. Am. Acad. Audiol.*, vol. 27, pp. 541–556, 2016.

[35] J. Cubick and T. Dau, "Validation of a virtual sound environment system for testing hearing aids," *Acta Acust united Ac*, vol. 102, 2016.

[36] F. Rumsey, "Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm," *J. Audio Eng. Soc.*, vol. 50, no. 9, pp. 651–666, 2002.

[37] T. H. Pedersen and N. Zacharov, "The development of a sound wheel for reproduced sound," *AES 138th Convention, Warsaw, Poland*, 2015.

[38] AES, "Practice for professional audio - subjective evaluation of loudspeakers." AES recommendation AES20-1996. Audio Engineering Society., 1996.

[39] N. Zacharov, T. Pedersen and C. Pike, "A common lexicon for spatial sound quality assessment - latest developments," *8th International Conference on Quality of Multimedia Experiemce QoMEX, Lisbon, Portugal*, 2016.

[40] A. Lindau, *Binaural Resynthesis of Acoustical Environments. Technology and Pereceptual Evaluation.* Phd thesis, Technische Universität Berlin, 2014.

[41] ITU-R, "Method for the subjective assessment of intermediate quality level of coding systems." Recommendation BS.1534-1. International Telecommunication Union, Geneva, 2003b.

[42] S. Zielinski, F. Rumsey and S. Bech, "On some biases encountered in modern audio quality listening tests - a review," *J. Audio Eng. Soc.*, vol. 56, no. 6, pp. 427–451, 2008.

[43] R. Mason, N. Ford, F. Rumsey and B. de Bruyn, "Verbal and non-verbal elicitation techniques in the subjective assessment of spatial sound reproduction," *AES 109th Convention, Los Angeles, USA*, 2000.

[44] N. Ford, F. Rumsey and B. de Bruyn, "Graphical elicitation techniques for subjective assessment of the spatial attributes of loudspeaker reproduction - a pilot investigation," *AES 110th Convention, Amsterdam, Netherlands*, 2001.

[45] H. Dillon, *Hearing Aids*. Thieme Medical Publishers, 2 ed., 2012.

[46] K. Klink, M. Schulte and M. Meis, "Measuring listening effort in the field of audiology - a literature review of methods, part 1," *Z. Audiol.*, vol. 51, no. 2, pp. 60–67, 2012.

[47] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *AES 108th Convention, Parma, Italy*, 2000.

[48] em32 Eigenmike microphone array release notes v17.0 by mh acoustics, "https://mhacoustics.com/sites/default/files/releasenotes.pdf," Sept. 2019.

[49] M. Brandner, *Sound Source Localization with the Eigenmike Microphone array - Evaluation and Analysis*. Project Thesis, Institut für elektronische Musik und Akustik, Graz, Austria, 2014.

[50] J. Skowronek, A. Raake, K. Hoeldtke and M. Geier, "Speech recordings for systematic assessment of multi-party conferencing," in *Forum Acusticum, Aalborg, Denmark*, 2011.

[51] J. Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the lombard reflex," *Speech Communication*, vol. 20, pp. 13–22, Nov. 1996.

[52] F. Zotter, H. Pomberger and M. Noisternig, "Energy-preserving ambisonic decoding," *Acta Acust united Ac*, vol. 98, pp. 37–47, 2012.

[53] I. T. Jolliffe, *Principal Component Analysis*. Springer Verlag, 2 ed., 2002.

[54] C. Kuhn-Rahloff, *Realitätstreue, Natürlichkeit, Plausibilität: Perzeptive Beurteilungen in der Elektroakustik*. Springer Verlag, 2012.

[55] A.W. Mills, "On the minimum audible angle," *J. Acoust. Soc. Am.*, vol. 30, no. 4, pp. 237–246, 1958.

[56] H. G. Hassager, T. May, A. Wiinberg and T. Dau, "Preserving spatial perception in rooms using direct-sound driven dynamic range compression," *J. Acoust. Soc. Am.*, vol. 141, no. 6, pp. 4556–4566, 2017.

[57] M. Yadav, D. Cabrera and W. L. Martens, "Auditory room size perceived from a room acoustic simulation with autophonic stimuli," *Acoustics Australia*, vol. 39, no. 3, pp. 101–105, 2011.

[58] M. Schutte, S.D. Ewert and L. Wiegrebe, "The percept of reverberation is not affected by visual room impression in virtual environments," *J. Acoust. Soc. Am.*, vol. 145, no. 3, 2019.

Appendix

Appendix A

The logged responses of the familiarization task of all subjects are listed as follows. Each table represents the responses to one change of scenes. The responses are displayed in their original language, German. The displayed labels are best fitting to summarize the responses in the eye of the investigator. It shall be noted that expressions like *Links/Rechts* (english: Left/Right) refer to the left and the right dialogues. A summary of all labels from these tables can be found in Table 4.1.

Table A.1 – Logged responses and the labels of the subjects in response to scene change **G1 to K1**.

Sub					Labels
1	Frau leicht näher gerückt	Alle bisschen trockener	Raum klingt kleiner, alle näher dran		Raumgrösse, Halligkeit, Distanz
2	Bisschen lauter	Hat sich was verändert, aber wenig			Lautheit
3	Tiefenverhältnisse sind anders	Frau weiter weg	Herren irgendwie anders angeordnet (in der Tiefe)		Tiefe, Distanz
4	Leichte Richtungsänderung nach links				Verteiltheit
5	Hallt weniger, trockener	Metallischer, klarer	kleinerer Raum		Halligkeit, Raumgrösse
6	Mehr Hall	Mehr Schall von hinten	räumlicher?	Klingen mehr von oben	Halligkeit
7	Wirkt ein bisschen breiter	Klingt näher dran	Klingt räumlicher	natürlicher, gewohnter (mehr wie echter Raum)	Breite, Distanz, Natürlichkeit
8	Weniger tieftönig	Weniger Hall			Klarheit, Halligkeit
9	Änderung wahrgenommen	Ggf ein Sprecher dezent umplaziert			Verteiltheit
10	Jetzt alles etwas tiefer	Er ewtas lauter	Nur drei Sprecher identifiziert		Tiefe, Lautheit
11	Ist breiter geworden	Weniger hallig als vorher			Breite, Halligkeit

Table A.2 – Logged responses and the labels of the subjects in response to scene change **K1 to K2**.

Sub					Labels
1	Mehr separiert	Rechts wirken etwas von oben (vgl. zu links)	Nicht so viel Hall wie in der 1.		Verteiltheit, Halligkeit
2	Dialoge jetzt aufgeteilt	Rechts eher oben	Links auch eher oben		Verteiltheit
3	Rechts und links separiert				Verteiltheit
4	Zwei Dialoge links, zwei rechts				Verteiltheit
5	Gespräche gesplittet rechts links	Etwas oben			Verteiltheit
6	Gespräche getrennt,	Einmal rechts	einmal links	Besser verständlich, weil getrennt	Verteiltheit, Höranstrengung
7	Gespräche aufgeteilt	Zwei links, zwei rechts			Verteiltheit
8	Rechts zwei	Links zwei wie vorher, leicht links	Dadurch klingt der Raum etwas grösser		Verteiltheit, Raumgrösse
9	Rechts und links zwei Gesprächspartner				Verteiltheit
10	Wie in die Szene reingerutscht	Rechts Mann & Frau	Links zwei Männer	Weniger hallig, weil gefühlt näher dran	Verteiltheit, Halligkeit
11	Gespräche nun getrennt	Raum gefühlt breiter geworden	Rechts ca 2 Uhr	Links ca auf 11 Uhr	Verteiltheit, Raumgrösse

Table A.3 – Logged responses and the labels of the subjects in response to scene change **K2 to G2**.

Sub					Labels
1	Immernoch separiert, gleiche Konstellationen	Rechts etwas höher gewandert	Mehr Hall bei allen	Rechts wirken recht nah aneinander	Halligkeit, Verteiltheit

2	Links höher als rechts	Eher minimer Unterschied	Rechts etwas dominanter als links		Lautheit
3	Rechts sind weiter nach hinten (winkel)	Links Winkel gleich	Tiefe schwer zu sagen		Tiefe
4	Änderung gehört, schwer zu beschreiben	Bisschen breiter geworden (eher wenig)			Breite
5	Halliger geworden (kleine Änderung)				Halligkeit
6	Rechts prominenter geworden	Rechts etwas weiter weg	Rechts hallt es etwas mehr	Links eher gleich geblieben	Distanz, Halligkeit
7	Rechtes Gespräch etwas lauter geworden	Gefühlt alles ein bisschen halliger	alles eher minim		Halligkeit, Lautheit
8	Minimale Änderung, bisschen mehr Hall				Halligkeit
9	Rechts vielleicht geändert, leicht anders plaziert	Rechts weiter unten (minim)			Verteiltheit
10	Vllt leichter Unterschied, aber nicht sicher				
11	Raum ist grösser geworden	Position gleich	Bisschen leiser geworden	Bisschen halliger	Raumgrösse, Halligkeit, Lautheit

Table A.4 – Logged responses and the labels of the subjects in response to scene change **G2 to G3**.

Sub					**Labels**
1	Links die beiden Sprecher separiert	Rechts auch separiert, Frau etwas links vom Mann			Verteiltheit
2	Links wie weiter weg	Rechts auch			Distanz
3	Keine Änderung gehört	Links vllt etwas präsenter			
4	Verhältnis zwischen den Sprechern geändert (Lautstärke)				Lautheit
5	nicht wahrgenommen				
6	Rechts etwas weiter weg, trotzdem noch gut verständlich	Links etwas mehr nach oben gewandert (aber noch nicht 90°)	Links auch langsam besser zu folgen	Links dezent lauter als rechts	Distanz, Lautheit, Höranstrengung
7	Keinen Unterschied gehört				
8	Rechts die Sprecher weiter nach Rechts				Verteiltheit
9	Links hat sich was verändert	Etwas an den Positionen verändert			Verteiltheit
10	Links vielleicht ein bisschen verrutscht				Verteiltheit
11	Wenn dann nur kleine Änderung	vllt minimale Lautheitsänderung			Lautheit

Table A.5 – Logged responses and the labels of the subjects in response to scene change **G3 to K3**.

Sub					**Labels**
1	Rechts beide etwas näher aneinander gerückt (aber nicht verschmolzen)	Links beide dezent runter gerutscht, aber noch separiert			Verteiltheit
2	Links einer ein bisschen separierter				Verteiltheit
3	Links einer fast weg, der andere sehr nah	Rechts Dame näher als Mann			Distanz
4	Änderungen gehört, DIstanzen etwas verändert				Distanz
5	Jetzt noch etwas halliger (vllt)				Halligkeit
6	Rechts etwas weiter weg, leicht weiter nach rechts	Links noch weiter oben	Links ist jetzt leichter zu folgen		Distanz, Höranstrengung

7	links ein Sprecher leiser geworden und nach links verschoben	links Quellbreite ist breiter			Lautheit, Verteiltheit, Breite
8	Rechts ein Stück weiter nach innen	Klang insgesamt etwas dumpfer, muffelig			Verteiltheit, Klarheit
9	Als ob Wände hinter den Sprechern aufgestellt wären	Effekt mehr links als rechts			Klarheit
10	Links näher gekommen	Weniger hallig, bisschen lauter	Rechts bisschen mehr im Hintergrund		Distanz, Halligkeit, Lautheit
11	Keine Änderung gehört				

Table A.6 – Logged responses and the labels of the subjects in response to scene change **K3 to K4**.

Sub					**Labels**
1	Rechts Sie ist weit weg!	Rechts klingen mehr auf einer Achse	Links beide weiter weg, einer etwas dumpfer	Bisschen mehr Hall insgesamt (rechts mehr)	Distanz, Halligkeit, Klarheit
2	Rechts weiter weg	Links jetzt dominanter, nicht so viel weiter weg			Distanz, Lautheit
3	Links weiter weg beide	Rechts Frau weiter weg			Distanz
4	Alle weiter weg	Lautstärke geändert			Distanz, Lautheit
5	Weiter weg, bisschen dumpfer, vllt etwas abgewandt	Rechts zieht mehr Aufmerksamkeit			Distanz, Zugewandtheit
6	Dumpfer geworden insgesamt, alle weiter auseinander	Links klingt als wären sie in nem anderen Raum	Links nuscheliger, ist schweiriger geworden	Rechts auch Raum gewechselt und weiter weg	Distanz, Verteiltheit, Klarheit
7	Links dumpfer geworden, weiter weg	Richtung links nicht mehr so eindeutig			Klarheit, Distanz
8	Rechts beide geändert, gleiche Richtung aber womöglich weiter weg	Rechts undeutlicher geworden, mehr Raum, mehr Nachhall	Eindruck links genauso		Distatnz, Klarheit, Raumgrösse, Halligkeit
9	Als ob die Wand jetzt zwischen VP und Sprechern	Rechts klingt als wären sie in anderem Raum	Oder als schauten sie von VP weg		Klarheit, Raumgrösse, Zugewandtheit, Distanz
10	Rechts weiter nach rechts				Verteiltheit
11	Rechts weiter weg	Links auch	Trotzdem noch gut verständlich		Distanz

Table A.7 – Logged responses and the labels of the subjects in response to scene change **K4 to G4**.

Sub					**Labels**
1	Mehr Hall insgesamt	Rechts Mann weniger hallig als Frau	Links zusammengerutscht, höher als Rechts	Links kleinerer Raum als rechts (vllt näher dran)	Halligkeit, Verteiltheit, Raumgrösse
2	Rechts beide noch weiter weg und weiter nach aussen	Links auch etwas weiter weg			Distanz
3	Rechts Herr präsenter (vllt näher gekommen)	Links bisschen näher (nur einer)			Distanz
4	Noch weiter weg, alles				Distanz
5	Weiter weg (?)	Raum ist grösser, halliger, bisschen leiser	Rechts ein bisschen unklarer		Distanz, Raumgrösse, Halligkeit, Lautheit, Klarheit
6	Links wieder Raum gewechselt, und weiter weg	Links weniger hallig, obwohl weiter weg	Rechts weiter weg, aber im gleichen Raum, leiser und schwerer verständlich	Links leichter zu folgen	Distanz, Halligkeit, Lautheit, Klarheit, Höranstrengung
7	Rechts weiter nach rechts & weiter weg	Links ein Sprecher extrem weit weg und schwer verständlich			Verteiltheit, Distanz, Höranstrengung
8	Klang insgesamt klarer, brillanter, weniger muffelig	Aber alle weiter weg	Klingt als gäbe es mehrere Räume für beide Gespräche	oder als wären alle in grossem Saal	Klarheit, Distanz, Raumgrösse
9	Alle leiser geworden,	Nicht weiter weg			Lautheit
10	Rechts weiter weg	Links auch leiser geworden	Halliger geworden		Distanz, Lautheit, Halligkeit

11	Alle noch weiter weg	Links vllt etwas in die mitte gerückt	Rechts vllt etwas weiter nach rechts	Alle dezent etwas höher	Distanz, Verteiltheit

Table A.8 – Logged responses and the labels of the subjects in response to scene change **G4 to G5**.

Sub					**Labels**
1	Alle weit weg	Links einer weniger deutlich (leiser) als der andere	Rechts Mann viel weiter weg als Frau		Klarheit, Distanz
2	Rechts Mann sehr weit weg, Frau näher	Links beide weiter weg			Distanz
3	Rechts Herr ist weit weg	Links beide weit weg	Links mglw geteilt		Distanz, Verteiltheit
4	Noch weiter weg, hallt viel mehr	Klingt fast als wären sie in anderem Raum			Distanz, Halligkeit
5	Weiter weg, dumpfer	Rechts Als wären sie in nem anderen Raum, um die Ecke gegangen	Links breiter geworden (hauptsächlich im Vergleich zu Anfang)		Distanz, Klarheit, Breite
6	Links noch weiter weg den Gang runter, bisschen nach rechts verschoben	Rechts auch etwas weiter nach rechts verschoben	Rechts leichter zu verstehen als links	Links klingt etwas weiter hoch, schwerer verständlich, dumpf, muffelig	Distanz, Verteiltheit, Höranstrengung, Klarheit
7	Alles weiter weg	Rechts vor allem	Links klingt fast wie aus nem Nebenraum	Links wie nicht mehr auf Listener gerichtet	Distanz, Zugewandtheit
8	Alle Quellen weiter weg	Beide weniger direkt, wie durch Wand getrennt	Richtungen immer noch gleich		Distanz, Klarheit
9	Links tönen als lägen sie im Lüftungsschacht	Rechts weit weg	Alles etwas grösser		Klarheit, Distanz, Raumgrösse
10	Links nochmal weiter weg	Alles weiter weg und leiser			Distanz, Lautheit
11	Noch halliger geworden	Rechts besser zu verstehen als links	Links dumpfer	Alles eher in grossem Raum	Halligkeit, Höranstrengung, Klarheit, Raumgrösse

Table A.9 – Logged responses and the labels of the subjects in response to scene change **G5 to K5**.

Sub					**Labels**
1	Alle wieder etwas näher dran	Raum wirkt kleiner, weniger Hall	Rechts er weiter weg als Sie		Distanz, Raumgrösse, Halligkeit
2	Alle jetzt ungefähr gleich weit weg	Rechts beide etwas getrennte (azimuth)	Links eher unauffälliger		Distanz, Verteiltheit
3	Links einer näher, der andere weiter weg	Links schwer zu erkennen, haben sich im Winkel geteilt	Rechts Frau näher, bisschen nach vorne gerückt	Rechts beide im Winkel geteilt	Distanz, Verteiltheit
4	Dumpf geworden	Klingt wie in nem Fass	Klingen etwas erhöht alle		Klarheit
5	Rechts getrennter (womöglich schon länger)	Rechts Frau klarer, zugewandt, er noch hinter der Ecke	Links einer redet klarer, der andere wie gegen die Wand		Verteiltheit, Klarheit
6	Links wieder näher gekommen, besser verständlich als die rechten	Links weniger muffelig, weniger hallig	Rechts Frau sehr dominant, im vgl zu eben nicht viel verändert		Distanz, Höranstrengung, Klarheit, Halligkeit
7	Deutlich weniger Nachhall	Näher dran, wie in einem trockenerem Raum	Anordnung der Sprecher gleich geblieben		Halligkeit, Distanz
8	Links beide näher gerückt	Insgesamt Raum kleiner geworden, niedrigere Decken			Distanz, Raumgrösse
9	Alle näher gekommen	Links als ob sie durch ein Medium sprechen (anderer Raum, Röhre), anders als eben	Rechts klingen mehr wie im gleichen Raum wie VP	Rechts Sie schaut zu VP, er schaut weg	Distanz, Klarheit, Zugewandtheit
10	Unsicher ob Änderung gehört	Ggf links noch etwas weiter weg			Distanz
11	Links wieder besser zu verstehen, weniger hallig	Rechts sind auch näher gekommen	Insgesamt bisschen trockener		Höranstrengung, Halligkeit, Distanz

Appendix B

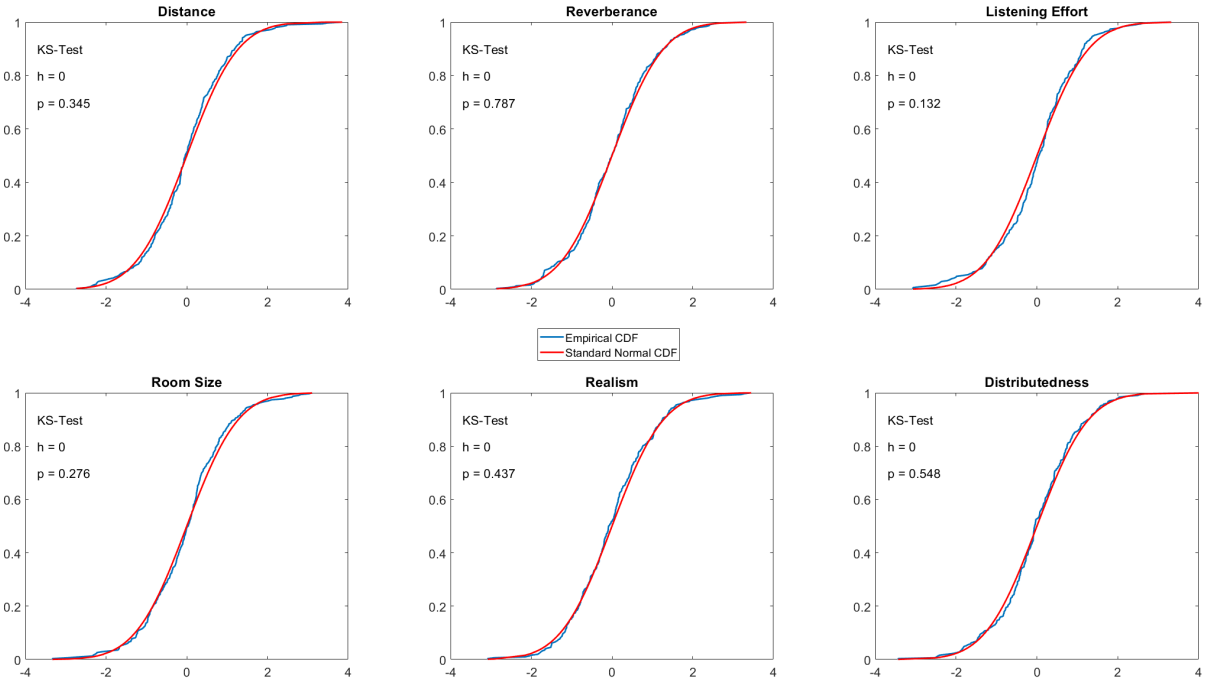


Figure B.1 – Display of cumulative distribution functions (CDF) of the centered and scaled residuals of the ANOVA models for each attribute versus the CDF of the standard normal distribution. The results from the Kolmogorov-Smirnov tests are displayed with their p-values. An h-value of 1 would reject the null-hypothesis that the data comes from a normal distribution at a 95 % confidence interval.

The output-tables of the multifactorial ANOVA from Matlab are listed as follows. The ANOVA was performed for each attribute separately with only two-way interactions to be displayed. The acronyms of the factors are Scene (SCN), Room (ROM), Subject (SUB) and Repetition (REP). Additionally displayed are numbers of factor levels (K), numbers of factor-level-combinations (n) and $\tilde{\delta}$ as a measure for effect size. The equation for $\tilde{\delta}$ is provided in Section 4.2, which is adapted from [6].

Table B.1 – Output-table of the multifactorial ANOVA from Matlab for the attribute **Distance**.

Source	Sum Sq.	d.f.	Singular?	Mean Sq.	F	Prob>F	K	n	$\tilde{\delta}$
SCN	168987.8	4	0	42246.95	260.86	1.15E-82	5	44	3.44
ROM	18706.9	1	0	18706.88	115.51	5.79E-22	2	110	1.45
SUB	3919.5	10	0	391.95	2.42	9.43E-03	11	20	0.49
REP	50.7	1	0	50.74	0.31	5.76E-01	2	110	0.08
SCN*ROM	5584.2	4	0	1396.04	8.62	1.75E-06	10	22	0.59
SCN*SUB	10970.5	40	0	274.26	1.69	9.36E-03	55	4	0.79
SCN*REP	688.8	4	0	172.19	1.06	3.76E-01	10	22	0.21
ROM*SUB	4016.9	10	0	401.69	2.48	7.79E-03	22	10	0.49
ROM*REP	2.5	1	0	2.55	0.02	9.00E-01	4	55	0.01
SUB*REP	2239.9	10	0	223.99	1.38	1.89E-01	22	10	0.36
Error	35953.4	222	0	161.95	[]	[]	[]	[]	[]
Total	257886.0	307	0	[]	[]	[]	[]	[]	[]

Table B.2 – Output-table of the multifactorial ANOVA from Matlab for the attribute **Reverberance**.

Source	Sum Sq.	d.f.	Singular?	Mean Sq.	F	Prob>F	K	n	$\tilde{\delta}$
SCN	84380.4	4	0	21095.09	83.99	2.54E-43	5	44	1.95
ROM	38942.8	1	0	38942.75	155.05	2.42E-27	2	110	1.68
SUB	9666.1	10	0	966.61	3.85	8.03E-05	11	20	0.62
REP	245.8	1	0	245.82	0.98	3.24E-01	2	110	0.13
SCN*ROM	4989.7	4	0	1247.44	4.97	7.47E-04	10	22	0.45
SCN*SUB	31097.5	40	0	777.44	3.10	6.17E-08	55	4	1.07
SCN*REP	909.1	4	0	227.28	0.90	4.62E-01	10	22	0.19
ROM*SUB	11561.9	10	0	1156.19	4.60	5.93E-06	22	10	0.66
ROM*REP	498.9	1	0	498.91	1.99	1.60E-01	4	55	0.16
SUB*REP	2177.3	10	0	217.73	0.87	5.65E-01	22	10	0.29
Error	55758.9	222	0	251.17	[]	[]	[]	[]	[]
Total	255516.7	307	0	[]	[]	[]	[]	[]	[]

Table B.3 – Output-table of the multifactorial ANOVA from Matlab for the attribute **Listening Effort**.

Source	Sum Sq.	d.f.	Singular?	Mean Sq.	F	Prob>F	K	n	$\tilde{\delta}$
SCN	88465.4	4	0	22116.36	104.48	6.78E-50	5	44	2.18
ROM	2638.9	1	0	2638.88	12.47	5.04E-04	2	110	0.48
SUB	29842.8	10	0	2984.28	14.10	3.31E-19	11	20	1.19
REP	115.4	1	0	115.38	0.55	4.61E-01	2	110	0.10
SCN*ROM	4048.5	4	0	1012.12	4.78	1.02E-03	10	22	0.44
SCN*SUB	46411.3	40	0	1160.28	5.48	4.72E-17	55	4	1.42
SCN*REP	3531.7	4	0	882.93	4.17	2.81E-03	10	22	0.41
ROM*SUB	2758.1	10	0	275.81	1.30	2.30E-01	22	10	0.35
ROM*REP	269.3	1	0	269.30	1.27	2.61E-01	4	55	0.12
SUB*REP	2773.5	10	0	277.35	1.31	2.26E-01	22	10	0.35
Error	46992.3	222	0	211.68	[]	[]	[]	[]	[]
Total	223097.2	307	0	[]	[]	[]	[]	[]	[]

Table B.4 – Output-table of the multifactorial ANOVA from Matlab for the attribute **Room Size**.

Source	Sum Sq.	d.f.	Singular?	Mean Sq.	F	Prob>F	K	n	$\tilde{\delta}$
SCN	99783.6	4	0	24945.91	119.30	3.79E-54	5	44	2.33
ROM	25608.7	1	0	25608.69	122.47	5.89E-23	2	110	1.49
SUB	7059.1	10	0	705.91	3.38	4.03E-04	11	20	0.58
REP	381.1	1	0	381.14	1.82	1.78E-01	2	110	0.18
SCN*ROM	2563.5	4	0	640.89	3.07	1.74E-02	10	22	0.35
SCN*SUB	17985.2	40	0	449.63	2.15	2.53E-04	55	4	0.89
SCN*REP	653.2	4	0	163.29	0.78	5.39E-01	10	22	0.18
ROM*SUB	6075.4	10	0	607.54	2.91	1.95E-03	22	10	0.53
ROM*REP	15.0	1	0	15.01	0.07	7.89E-01	4	55	0.03
SUB*REP	1204.1	10	0	120.41	0.58	8.33E-01	22	10	0.23
Error	46419.1	222	0	209.10	[]	[]	[]	[]	[]
Total	217217.4	307	0	[]	[]	[]	[]	[]	[]

Table B.5 – Output-table of the multifactorial ANOVA from Matlab for the attribute **Realism**.

Source	Sum Sq.	d.f.	Singular?	Mean Sq.	F	Prob>F	K	n	$\tilde{\delta}$
SCN	40046.9	4	0	10011.71	22.10	2.27E-15	5	44	1.00
ROM	1967.8	1	0	1967.80	4.34	3.83E-02	2	110	0.28

SUB	13197.4	10	0	1319.74	2.91	1.90E-03	11	20	0.54
REP	1653.8	1	0	1653.75	3.65	5.74E-02	2	110	0.26
SCN*ROM	1535.1	4	0	383.78	0.85	4.97E-01	10	22	0.19
SCN*SUB	57265.8	40	0	1431.65	3.16	3.44E-08	55	4	1.08
SCN*REP	854.6	4	0	213.65	0.47	7.57E-01	10	22	0.14
ROM*SUB	7757.4	10	0	775.74	1.71	7.92E-02	22	10	0.40
ROM*REP	14.6	1	0	14.57	0.03	8.58E-01	4	55	0.02
SUB*REP	1494.1	10	0	149.41	0.33	9.72E-01	22	10	0.18
Error	100578.9	222	0	453.06	[]	[]	[]	[]	[]
Total	223249.6	307	0	[]	[]	[]	[]	[]	[]

Table B.6 – Output-table of the multifactorial ANOVA from Matlab for the attribute **Distributedness**.

Source	Sum Sq.	d.f.	Singular?	Mean Sq.	F	Prob>F	K	n	$\tilde{\delta}$
SCN	199246.2	4	0	49811.55	460.49	3.26E-106	5	44	4.58
ROM	3722.9	1	0	3722.89	34.42	1.60E-08	2	110	0.79
SUB	5694.1	10	0	569.41	5.26	6.03E-07	11	20	0.73
REP	172.9	1	0	172.94	1.60	2.07E-01	2	110	0.17
SCN*ROM	4958.1	4	0	1239.54	11.46	1.79E-08	10	22	0.68
SCN*SUB	21033.3	40	0	525.83	4.86	9.14E-15	55	4	1.34
SCN*REP	1108.8	4	0	277.19	2.56	3.93E-02	10	22	0.32
ROM*SUB	1720.7	10	0	172.07	1.59	1.11E-01	22	10	0.39
ROM*REP	578.0	1	0	577.98	5.34	2.17E-02	4	55	0.25
SUB*REP	2142.7	10	0	214.27	1.98	3.65E-02	22	10	0.43
Error	24013.7	222	0	108.17	[]	[]	[]	[]	[]
Total	267900.0	307	0	[]	[]	[]	[]	[]	[]

Table B.7 – Output table of post-hoc multiple comparison tests (Tuckey HSD) for the factors SCN and ROM on the ratings of **Distance**. The values indicate the differences in the mean ratings, colour indicates significance (□ $p < 0.001$ / □ $p < 0.01$ / □ $p < 0.05$ / □ n.s.). The indices stand for rooms Gravel and Kircher and scenes #1 to #5. The bold values emphasize the comparisons of the same scenes between the two rooms.

	G1	G2	G3	G4	G5	K1	K2	K3	K4	K5
G1	0	2.1	-2.5	-32.4	-55.3	**21.3**	1.3	22	-19	-31.4
G2		0	-4.6	-34.5	-57.4	19.3	**-0.8**	20	-21	-33.5
G3			0	-29.9	-52.8	23.9	3.8	**24.5**	-16.5	-28.9
G4				0	-22.9	53.8	33.7	54.5	**13.5**	1
G5					0	76.7	56.6	77.3	36.3	**23.9**
K1						0	-20	0.7	-40.3	-52.8
K2							0	20.7	-20.3	-32.7
K3								0	-41	-53.4
K4									0	-12.4
K5										0

Table B.8 – Output table of post-hoc multiple comparison tests (Tuckey HSD) for the factors SCN and ROM on the ratings of **Reverberance**. The values indicate the differences in the mean ratings, colour indicates significance (□ $p < 0.001$ / □ $p < 0.01$ / □ $p < 0.05$ / □ n.s.). The indices stand for rooms Gravel and Kircher and scenes #1 to #5. The bold values emphasize the comparisons of the same scenes between the two rooms.

	G1	G2	G3	G4	G5	K1	K2	K3	K4	K5
G1	0	11.4	10	-15	-37.1	**28.9**	31.5	28	0.3	-0.5
G2		0	-1.4	-26.5	-48.5	17.5	**20.1**	16.6	-11.1	-11.9
G3			0	-25.1	-47.2	18.9	21.5	**18**	-9.7	-10.5
G4				0	-22.1	44	46.5	43.1	**15.4**	14.5
G5					0	66	68.6	65.2	37.4	**36.6**
K1						0	2.6	-0.9	-28.6	-29.4
K2							0	-3.5	-31.2	-32
K3								0	-27.7	-28.5
K4									0	-0.8
K5										0

Table B.9 – Output table of post-hoc multiple comparison tests (Tuckey HSD) for the factors SCN and ROM on the ratings of **Listening Effort**. The values indicate the differences in the mean ratings, colour indicates significance (□ $p < 0.001$ / □ $p < 0.01$ / □ $p < 0.05$ / □ n.s.). The indices stand for rooms Gravel and Kircher and scenes #1 to #5. The bold values emphasize the comparisons of the same scenes between the two rooms.

	G1	G2	G3	G4	G5	K1	K2	K3	K4	K5
G1	0	35.8	34	14.2	-12.9	**0.2**	34.7	38.2	25.2	3.8
G2		0	-1.8	-21.5	-48.6	-35.6	**-1**	2.4	-10.5	-32
G3			0	-19.8	-46.9	-33.8	0.7	**4.2**	-8.8	-30.2
G4				0	-27.1	-14	20.5	24	**11**	-10.4
G5					0	13	47.6	51	38.1	**16.7**
K1						0	34.5	38	25	3.6
K2							0	3.5	-9.5	-30.9
K3								0	-13	-34.4
K4									0	-21.4
K5										0

Table B.10 – Output table of post-hoc multiple comparison tests (Tuckey HSD) for the factors SCN and ROM on the ratings of **Room Size**. The values indicate the differences in the mean ratings, colour indicates significance (□ $p < 0.001$ / □ $p < 0.01$ / □ $p < 0.05$ / □ n.s.). The indices stand for rooms Gravel and Kircher and scenes #1 to #5. The bold values emphasize the comparisons of the same scenes between the two rooms.

	G1	G2	G3	G4	G5	K1	K2	K3	K4	K5
G1	0	8.2	-1.2	-21.1	-41.6	**24.2**	16.5	19.9	-3.5	-16.2
G2		0	-9.4	-29.3	-49.8	16	**8.3**	11.7	-11.7	-24.4
G3			0	-19.9	-40.4	25.4	17.7	**21.1**	-2.3	-15
G4				0	-20.5	45.3	37.5	41	**17.5**	4.9
G5					0	65.8	58.1	61.5	38.1	**25.4**
K1						0	-7.7	-4.3	-27.7	-40.3
K2							0	3.4	-20	-32.6
K3								0	-23.4	-36
K4									0	-12.6
K5										0

Table B.11 – Output table of post-hoc multiple comparison tests (Tuckey HSD) for the factors SCN and ROM on the ratings of **Realism**. The values indicate the differences in the mean ratings, colour indicates significance (□ $p < 0.001$ / □ $p < 0.01$ / □ $p < 0.05$ / □ n.s.). The indices stand for rooms Gravel and Kircher and scenes #1 to #5. The bold values emphasize the comparisons of the same scenes between the two rooms.

	G1	G2	G3	G4	G5	K1	K2	K3	K4	K5
G1	0	-29.1	-32.8	-16.8	-3.1	**-6.6**	-33.6	-30.7	-22.6	-15.1
G2		0	-3.7	12.3	25.9	22.5	**-4.5**	-1.6	6.5	13.9
G3			0	16	29.7	26.2	-0.8	**2.1**	10.2	17.7
G4				0	13.6	10.2	-16.8	-13.9	**-5.8**	1.6
G5					0	-3.4	-30.4	-27.5	-19.4	**-12**
K1						0	-27	-24.1	-16	-8.6
K2							0	2.9	11	18.4
K3								0	8.1	15.5
K4									0	7.4
K5										0

Table B.12 – Output table of post-hoc multiple comparison tests (Tuckey HSD) for the factors SCN and ROM on the ratings of **Distributedness**. The values indicate the differences in the mean ratings, colour indicates significance (□ $p < 0.001$ / □ $p < 0.01$ / □ $p < 0.05$ / □ n.s.). The indices stand for rooms Gravel and Kircher and scenes #1 to #5. The bold values emphasize the comparisons of the same scenes between the two rooms.

	G1	G2	G3	G4	G5	K1	K2	K3	K4	K5
G1	0	-35.4	-36.1	-53.7	-69	**10.1**	-37.5	-35.8	-45.4	-48.8
G2		0	-0.7	-18.4	-33.6	45.5	**-2.1**	-0.5	-10	-13.4
G3			0	-17.6	-32.9	46.2	-1.4	**0.3**	-9.3	-12.7
G4				0	-15.2	63.8	16.2	17.9	**8.4**	5
G5					0	79.1	31.5	33.1	23.6	**20.2**
K1						0	-47.6	-45.9	-55.5	-58.9
K2							0	1.7	-7.9	-11.3
K3								0	-9.5	-13
K4									0	-3.4
K5										0

Appendix C

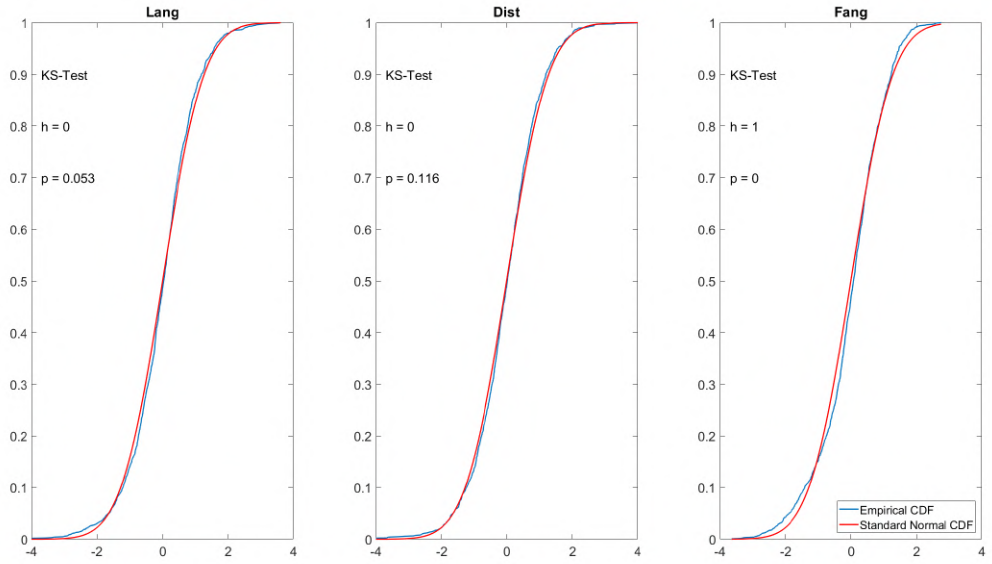


Figure C.1 – Display of cumulative distribution functions (CDF) of the centered and scaled residuals of the ANOVA models for each outcome variable (Distance (Dist), Lateral Angle (Lang) and Facing Angle (Fang)) versus the CDF of the standard normal distribution. The results from the Kolmogorov-Smirnov tests are displayed with their p-values. An h-value of 1 would reject the null-hypothesis that the data comes from a normal distribution at a 95 % confidence interval. The residuals for the Fang are not normally distributed, consequently they will not be analysed in the ANOVA.

The output-tables of the multifactorial ANOVA from Matlab are listed as follows. The ANOVA was performed for distance and lateral angle of the figurine placements separately with only two-way interactions to be displayed. The acronyms of the factors are Scene (SCN), Room (ROM), Subject (SUB) and Repetition (REP). Additionally to the Matlab output, numbers of factor levels (K), numbers of factor-level-combinations (n) and $\tilde{\delta}$ as a measure for effect size are displayed. The equation for $\tilde{\delta}$ is provided in subsection 4.2, which is adapted from [6].

Table C.1 – Output table of the multifactorial ANOVA from Matlab for the Lateral Angle placements of the figurine task. The tables are extended by K as factor levels, n as factor-level combinations and $\tilde{\delta}$ as a measure for effect size.

Source	Sum Sq.	d.f.	Singular?	Mean Sq.	F	Prob>F	K	n	$\tilde{\delta}$
SPK	278.1	3	0	92.70	248.40	6.2E-123	4	220	1.50
SCN	22.1	4	0	5.53	14.83	8.3E-12	5	176	0.41
ROM	16.1	1	0	16.08	43.08	8.1E-11	2	440	0.44
SUB	127.9	10	0	12.79	34.26	2.3E-58	11	80	0.93
REP	0.3	1	0	0.31	0.84	3.6E-01	2	440	0.06
SPK*SCN	102.7	12	0	8.56	22.93	6.9E-46	20	44	0.81
SPK*ROM	0.3	3	0	0.10	0.27	8.5E-01	8	110	0.05
SPK*SUB	84.0	30	0	2.80	7.51	2.1E-28	44	20	0.72
SPK*REP	0.3	3	0	0.09	0.24	8.7E-01	8	110	0.04
SCN*ROM	7.1	4	0	1.79	4.79	7.9E-04	10	88	0.22
SCN*SUB	81.5	40	0	2.04	5.46	1.3E-23	55	16	0.71
SCN*REP	1.3	4	0	0.33	0.90	4.7E-01	10	88	0.10

ROM*SUB	4.7	10	0	0.47	1.27	2.4E-01	22	40	0.17
ROM*REP	0.1	1	0	0.15	0.39	5.3E-01	4	220	0.03
SUB*REP	18.3	10	0	1.83	4.90	5.8E-07	22	40	0.34
Error	408.7	1095	0	0.37	[]	[]	[]	[]	[]
Total	1231.0	1231	0	[]	[]	[]	[]	[]	[]

Table C.2 – Output table of the multifactorial ANOVA from Matlab for the Distance placements of the figurine task. The tables are extended by K as factor levels, n as factor-level combinations and $\tilde{\delta}$ as a measure for effect size.

Source	Sum Sq.	d.f.	Singular?	Mean Sq.	F	Prob>F	K	n	$\tilde{\delta}$
SPK	195.2	3	0	65.08	207.21	2.0E-106	4	220	1.37
SCN	51.4	4	0	12.85	40.90	5.6E-32	5	176	0.68
ROM	46.8	1	0	46.79	148.97	3.2E-32	2	440	0.82
SUB	202.6	10	0	20.26	64.49	5.7E-103	11	80	1.27
REP	0.5	1	0	0.50	1.61	2.1E-01	2	440	0.09
SPK*SCN	122.9	12	0	10.24	32.61	1.2E-64	20	44	0.97
SPK*ROM	18.1	3	0	6.05	19.25	3.6E-12	8	110	0.39
SPK*SUB	56.7	30	0	1.89	6.02	2.3E-21	44	20	0.65
SPK*REP	1.5	3	0	0.50	1.58	1.9E-01	8	110	0.11
SCN*ROM	5.5	4	0	1.37	4.38	1.6E-03	10	88	0.21
SCN*SUB	70.7	40	0	1.77	5.63	1.3E-24	55	16	0.72
SCN*REP	1.6	4	0	0.40	1.27	2.8E-01	10	88	0.11
ROM*SUB	14.6	10	0	1.46	4.65	1.6E-06	22	40	0.33
ROM*REP	0.2	1	0	0.18	0.58	4.5E-01	4	220	0.04
SUB*REP	35.8	10	0	3.58	11.41	8.9E-19	22	40	0.52
Error	343.9	1095	0	0.31	[]	[]	[]	[]	[]
Total	1231.0	1231	0	[]	[]	[]	[]	[]	[]

Table C.3 – Output table of post-hoc multiple comparison tests (Tuckey HSD) for the factor SCN for Speaker 1 separately on the Lateral Angle in [°] and Distance in pixel units placement of the figurine alignment task. The values show the differences in the means, colour indicates significance (□ $p < 0.001$ / □ $p < 0.01$ / □ $p < 0.05$ / □ n.s.).

	Lateral Angle					Distance				
SPK1	SCN1	SCN2	SCN3	SCN4	SCN5	SCN1	SCN2	SCN3	SCN4	SCN5
SCN1	0	-5.1	-3.6	-1.4	0.3	0	8.9	5.1	-13.9	-85.1
SCN2		0	1.5	3.7	5.4		0	-3.8	-22.8	-94
SCN3			0	2.2	3.9			0	-19	-90.2
SCN4				0	1.7				0	-71.2
SCN5					0					0

Table C.4 – Output table of post-hoc multiple comparison tests (Tuckey HSD) for the factor SCN for Speaker 2 separately on the Lateral Angle in [°] and Distance in pixel units placement of the figurine alignment task. The values show the differences in the means, colour indicates significance (□ $p < 0.001$ / □ $p < 0.01$ / □ $p < 0.05$ / □ n.s.).

	Lateral Angle					Distance				
SPK2	SCN1	SCN2	SCN3	SCN4	SCN5	SCN1	SCN2	SCN3	SCN4	SCN5
SCN1	0	-0.9	3.3	4.4	1.4	0	5.7	16.3	-24.5	-66.7
SCN2		0	4.2	5.3	2.3		0	10.6	-30.2	-72.4
SCN3			0	1.1	-1.9			0	-40.9	-83
SCN4				0	-3				0	-42.2
SCN5					0					0

Table C.5 – Output table of post-hoc multiple comparison tests (Tuckey HSD) for the factor SCN for Speaker 3 separately on the Lateral Angle in [°] and Distance in pixel units placement of the figurine alignment task. The values show the differences in the means, colour indicates significance (□ $p < 0.001$ / □ $p < 0.01$ / □ $p < 0.05$ / □ n.s.).

	Lateral Angle					Distance				
SPK3	SCN1	SCN2	SCN3	SCN4	SCN5	SCN1	SCN2	SCN3	SCN4	SCN5
SCN1	0	68.4	59.5	64.1	69	0	-49.4	-33.7	-155	-211.7
SCN2		0	-8.9	-4.3	0.6		0	15.7	-105.6	-162.3
SCN3			0	4.5	9.5			0	-121.3	-177.9
SCN4				0	4.9				0	-56.7
SCN5					0					0

Table C.6 – Output table of post-hoc multiple comparison tests (Tuckey HSD) for the factor SCN for Speaker 4 separately on the Lateral Angle in [°] and Distance in pixel units placement of the figurine alignment task. The values show the differences in the means, colour indicates significance (□ $p < 0.001$ / □ $p < 0.01$ / □ $p < 0.05$ / □ n.s.).

	Lateral Angle					Distance				
SPK4	SCN1	SCN2	SCN3	SCN4	SCN5	SCN1	SCN2	SCN3	SCN4	SCN5
SCN1	0	71.9	67.6	67.5	70.5	0	-52.1	-47.8	-155.3	-226.2
SCN2		0	-4.3	-4.4	-1.4		0	4.3	-103.2	-174.1
SCN3			0	-0.1	2.9			0	-107.5	-178.4
SCN4				0	3				0	-70.9
SCN5					0					0

Table C.7 – Output table of post-hoc multiple comparison tests (Tuckey HSD) for the factors SCN and ROM on the results of Lateral Angle placement. The values show the differences in the means in [°], colour indicates significance (□ $p < 0.001$ / □ $p < 0.01$ / □ $p < 0.05$ / □ n.s.) The indices stand for rooms Gravel and Kircher and scenes #1 to #5. The bold values indicate differences of the same scenes between the rooms.

	G1	G2	G3	G4	G5	K1	K2	K3	K4	K5
G1	0	35	33.6	36.3	34.9	**-1.4**	30.8	28.5	29.6	34.3
G2		0	-1.4	1.3	-0.1	-36.4	**-4.2**	-6.5	-5.4	-0.7
G3			0	2.8	1.4	-34.9	-2.7	**-5.1**	-3.9	0.8
G4				0	-1.4	-37.7	-5.5	-7.8	**-6.7**	-2
G5					0	-36.3	-4.1	-6.5	-5.3	**-0.6**
K1						0	32.2	29.8	31	35.7
K2							0	-2.3	-1.2	3.5
K3								0	1.2	5.8
K4									0	4.7
K5										0

Table C.8 – Output table of post-hoc multiple comparison tests (Tuckey HSD) for the factors SCN and ROM on the results of Distance placement. The values show the differences in the means in pixel units, colour indicates significance (□ $p < 0.001$ / □ $p < 0.01$ / □ $p < 0.05$ / □ n.s.) The indices stand for rooms Gravel and Kircher and scenes #1 to #5. The bold values indicate differences of the same scenes between the rooms.

	G1	G2	G3	G4	G5	K1	K2	K3	K4	K5
G1	0	-18.3	-22	-102.3	-161.4	**31.8**	6.6	23.8	-40.3	-101.6
G2		0	-3.7	-84	-143.1	50.1	**24.9**	42	-22	-83.3
G2			0	-80.3	-139.4	53.8	28.7	**45.8**	-18.3	-79.6
G4				0	-59.1	134.1	108.9	126.1	**62**	0.7
G5					0	193.2	168	185.2	121.1	**59.8**
K1						0	-25.2	-8	-72.1	-133.4
K2							0	17.1	-46.9	-108.3
K3								0	-64	-125.4
K4									0	-61.3
K5										0

Appendix D

Table D.1 – Logged descriptions of the two presented rooms for each subject. Additionally, the guessed rooms corresponding to the pictures in Figure 3.2 are displayed in columns G and K. The ground truth is G = 2 and K = 4. The order of presentation is displayed in the last column. The descriptions are displayed in German.

SUB	Room G	Room K	G	K	Order
1	Grosser Raum; Glatte Wände; Holzboden, glatt; Leerer Wohnraum; Nicht viele Fenster;	Eher klein; Wenig Hall, nichts hallt wirklich nachGedämpft, Teppich; Tiefere Decke; Länger, eher Schuhkartonförmig;	3	2	GK
2	Weniger hallig als B; Aber auch gross; Wie Grossraumbüro; Leicht hallig (vorallem Mann Rechts); Mehr Inhalt, wie Büro Raum nicht regulär, mglw. verwinkelt; Oder Trennwände;	Grösserer Raum; Eher halliger, wenig Inhalt, viele Reflexionen; Nicht ganz Kirche, mehr wie eine Halle, vllt eine Arbeitshalle;	3	2	KG
3	Grosser Raum; Sprecher weit wegRecht hallig; Raum ist leer;	Um einiges kleiner; Reflexionen, auch an Gegenständen, TischeMittelgrosser; Meetingraum;	3	2	GK
4	Noch grösser wie B; Hallt mehr als B; Auch wenig Inhalt; Sonst relativ ähnlich wie; B	Relativ grosser Raum, 2-3* grösser wie Messraum; Hohe Decken; Boden Parkett, eher glatt; Wenig Inhalt; (Möbel);	3	4	KG
5	Hallig, glatte Flächen Grösser als B; Länglich, rechteckig; Rechts stehen in ner EckeRecht wenig Gegenstände Wie Maschinenhalle; Eher hoch;	Im Raum Abtrennwände zwischen Sprechern-Kleinerer Raum; Eher glatte Flächen; Wie in nem Labor, mit Gegenständen; drin;	3	2	GK
6	Raum grösser; Halliger als B; Nicht ganz wie Kirche, aber schlimm; Klingt wie zwei verschiedene Räume für beide Gespräche (rechts wie in anderem Raum)Links viel; präsenter; Rechts kommt wie aus Hallraum; Links könnte auch aus Labor kommen;	Ein Raum; Eher kleiner; Muffelig, dumpf; Kein Messraum, akustisch nicht; bearbeitetViel Reflexion, nicht Hallraum, aber vielTische im Raum?; Meetingraum bei uns; Szene wie kurz vor Meeting Beginn; Vllt Klassenzimmer	2	1	KG
7	Hallig wie Foyer; Grösser; Vergleich mit Bekanntem Raum aus Schulzeiten; (Foyer der Schule) kastenförmig; Keine hohe Decke; Mittelmässig bis schlecht gedämmt;	Stärker gedämmt (Teppich); Assoziation mit Lehrerzimmer (seiner Schule)Mit Tischen drin; Raum verzweigter (mit um die Ecke); Auf jeden Fall kleiner;	3	4	GK

8	Deutlich halliger; Grosser Raum; Hohe Decke; Könnte eine Turnhalle sein; Glatte Flächen;	Niedrige Decke; Eher gross, wie SeminarraumRelativ viel; Dämmung;	3	2	KG
9	Sehr hallig; Rechts stehen mehr im freien; Links mehr an der Wand; Nicht viel Gegenstände drin; Raum könnte leer sein; Referat-Zimmer (vgl. mit Raum von der ETH);	Weniger hallig; Mglw. Absorptionsflächen drin; Kleinerer Raum; Nicht überall gleiche Absorption; Unterschiedlich ausgekleidete Wände; Wand näher an den Rechten;	3	2	GK
10	Grosser Raum; Hohe Decke; Stark reflektierend; Grosser Keller, vllt TiefgarageWenig Inhalt; Kein Teppich;	Weniger Hallig; Weniger Keller-likeKleiner Raum; Teppich drin; Mehr Inhalt;	3	2	KG
11	Grösser als B; Könnte eine Kirche sein; Auch sehr hallig; Keine akustische Bearbeitung;	Grosser Raum; Hallig; Keine Kirche; Glatte Wände, keine akustische Bearbeitung; Runder Raum;	3	4	GK