The final publication is available at Springer via http://dx.doi.org/10.1007/s10055-015-0270-8
Schoeffler, M., Gernert, J.L., Neumayer, M. et al. Virtual Reality (2015) 19: 181. doi:10.1007/s10055-015-0270-8

On the Validity of Virtual Reality-based Auditory Experiments: A Case Study about Ratings of the Overall Listening Experience

Michael Schoeffler · Jan Lukas Gernert · Maximilian Neumayer · Susanne Westphal · Jürgen Herre

Abstract In recent years, new developments have led to an increasing number of Virtual Reality-based experiments, but little is known about their validity compared to real-world experiments. To this end, an experiment was carried out which compares responses given in a real-world environment to responses given in a Virtual Reality (VR) environment. In the experiment, thirty participants rated the overall listening experience of music excerpts while sitting in a cinema and a listening booth being in a real-world environment and in a VR environment. In addition, the VR system that was used to carry out the sessions in the VR environment is presented in detail.
Results indicate that there are only minor statistically significant differences between the two environments when the overall listening experience is rated. Furthermore, in the real-world environment, the ratings given in the listening booth were slightly higher than in the cinema.

Keywords Virtual Reality-based Experiments, Overall Listening Experience, Convolution Engine, Oculus Rift

Michael Schoeffler
International Audio Laboratories Erlangen
A joint institution of Fraunhofer IIS and Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
Am Wolfsmantel 33, 91058 Erlangen, Germany
Tel.: +49 9131 85-20515
E-mail: michael.schoeffler@audiolabs-erlangen.de

Jan Lukas Gernert, Maximilian Neumayer, Susanne Westphal, Jürgen Herre
International Audio Laboratories Erlangen
A joint institution of Fraunhofer IIS and Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
Am Wolfsmantel 33, 91058 Erlangen, Germany

1 Introduction

The ways auditory experiments are conducted is evolving. Until a few years ago, the majority of experiments has mainly been conducted in laboratory environments. The advantage of laboratory experiments is that confounding variables can be controlled to a certain extend and experimenters can monitor the participants.

Nowadays, conducting so-called web-based experiments (or Internet experiments) is becoming more popular (Welch and Krantz, 1996; Schoeffler et al., 2015; Pysiewicz, 2014). The validity of web-based experiments has been evaluated by various researchers who came to the conclusion that this type of experiment results in very similar outcome compared to real-world experiments[1] (Pysiewicz, 2014; Schoeffler et al., 2013b). Compared to laboratory experiments, web-based experiments support a simplified recruitment process since the experiment can be accessed by any web-enabled device from all over the world. One of the main disadvantages of web-based experiments is that they lack full control over the experiment procedure. Moreover, web-based experiments are not suited for all types of experiments related to spatial audio, especially, e. g., if they require specific room acoustics. VR experiments have the potential to overcome this issue by providing an authentic visual and auditory representation of a specific room.

The term VR describes the simulation of an environment that creates the immersion to be present in places in the real or in an imagined world (Steuer, 1992). A simulated physical presence becomes only plausible

[1] In this paper, the term *real-world experiment* is used to describe experiments that were conducted under laboratory conditions in the real-world (and not in the VR). Sometimes this term is also used as a synonym for the term *field experiment*.

for humans when several requirements are met. Stanney (1995) described issues related to these requirements which must be solved to reach the full potential of VR systems, including human performance, user characteristics, visual and auditory perception, absence of cybersickness and social impacts (see also Stanney et al. (1998)). Some of these issues were investigated by experimental studies in order to measure their influence on the perceived level of presence (van Dam et al., 2002; Sanchez-Vives and Slater, 2005; Schuemie et al., 2001). One could argue that an authentic visual stimulus might not be important when conducting auditory experiments, but many studies have shown that the visual stimulus has a significant influence when participants give responses to auditory stimuli (Seeber and Fastl, 2004; Werner and Siegel, 2011; Werner et al., 2012; Gorzel et al., 2012).

Increasing computational power and capabilities of VR devices, e. g. higher display resolution of head-mounted displays, allow rendering virtual environments more accurately in real-time, resulting in more authentic visual and auditory representations. An authentic auditory representation of an environment can be achieved by binaural synthesis. When using binaural synthesis, audio algorithms process two signals, where each signal corresponds to a single ear of a listener. For example, room impulse responses of a room (which contain information about the acoustical characteristics of a room) can be measured by a dummy head having one microphone in each ear. A set of such room impulse responses is called binaural room impulse responses (BRIRs). When convolving a (single-channel) auditory stimulus with a BRIR, the resulting audio signal sounds like an accurate reproduction of the stimulus played back in the measured room at the exact position where the BRIRs were recorded.

In the future, web-based experiments might be combined with VR experiments having both advantages: theoretical access to millions of participants and the possibility to create any type of environment, provided participants have a compatible VR device at home. In order to draw scientifically correct conclusions from such experiments, both types, web-based and VR experiments, must be valid compared to real-world experiments. Therefore, results of VR experiments must be compared to the results of real-world experiments, to find out under which circumstances they become valid.

This work investigates the validity of an auditory experiment conducted in a real-world and a VR environment. In Section 3, a system is presented which was developed to carry out the VR sessions of the auditory experiment. The system consists of a modified game engine which allows a virtual scene to be rendered on a head-mounted display, an apparatus to measure BRIRs for any number of head positions, and a real-time convolution engine for convolving the BRIRs with auditory stimuli. A pair of headphones is used to reproduce the convolved auditory stimuli.

In the experiment, participants rated their perceived overall listening experience (OLE) while listening to short music excerpts in different types of rooms. OLE is a term used for describing the sensation, perception, and cognition that is active when someone listens to sound events. A rating of the OLE reflects the resulting enjoyment of listening to this sound event (Schoeffler and Herre, 2013, 2014a). Thus, when listeners rate the OLE, they are asked to take into account every factor that influences their enjoyment while listening to something. Such factors of influence might include song, lyrics, audio quality, listener's mood, listening room, and reproduction system. Since so many factors might have an influence, the OLE has been shown to be a very holistic attribute, where each person has different preferences (Schoeffler and Herre, 2014b). As the term OLE is only used in the context of listening experience, it is a subattribute of the more general term Quality of Experience which describes the degree of enjoyment or satisfaction of humans while using a system. Le Callet et al. (2012) defined Quality of Experience as follows:

> "Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state."

The participants of the presented experiment evaluated the OLE of short music excerpts in a cinema and in a small listening booth. The short music excerpts were mixed in 5.1 surround sound. In order to contribute to the validation of VR experiments, the participants were present in a real cinema and listening booth, and also in a virtual representation of the two rooms. Two main hypotheses are stated that were answered by means of the presented experiment:

Hypothesis I
: The OLE ratings given in the VR environment do not differ from those given in the real-world experiment.

Hypothesis II
: The OLE ratings that were given while being present in the (real-world) cinema do not differ from those given in the (real-world) listening booth.

Besides adding questions targeting those hypotheses to the experiment, we also added some questions that are related to the acoustical characteristics of the room and reproduction system used. For example, we asked the participants how far away they perceived a sound source or what amount of reverberation they estimated. The goal of asking these questions is not to find out whether acoustical characteristics are perceived differently in a VR experiment and in a real-world experiment. In our opinion, answering such a research question would require a dedicated and more comprehensive experiment. Instead, we want to give some indications as to which might be the basis for hypotheses of experiments conducted in the near future.

The detailed methodology of the experiment is described in Section 4.1. By comparing the responses given in the real-world environment and given in the VR environment, we give some indication to what degree the VR environment was valid compared to the real-world environment (see Section 4.2). The results are subsequently discussed in Section 4.3.

In the next section, we give a short introduction into three approaches used to create a VR environment. Furthermore, state-of-the-art VR systems are presented focusing on the audio rendering. Next, studies are reviewed that already compared the results of a VR experiment to results of a real-world experiment. Lastly, we summarize the current literature that is related to our second hypothesis and might give some hints as to how much the room influences the OLE.

2 Related Work

2.1 Virtual Reality Technologies

VR environments can be created by various approaches. The most common ones are cave automatic virtual environments (CAVEs) and head-mounted displays (HMDs).

In a CAVE, images of a scene are projected on multiple walls (and sometimes also on the floor and ceiling) of a room (Cruz-Neira et al., 1992). The images of the scene are updated according to the user's head position and viewed by stereoscopic glasses. One major advantage of a CAVE is that the field of view is typically very wide, allowing the user to walk around within the virtual scene without decreasing the state of immersion. A disadvantage of CAVEs is that a lot of equipment (projectors, head-tracker, glasses etc.) and space are needed to create a VR environment. Both a wide field of view and the need for no additional resources is offered by binocular HMDs which are devices, worn on the head, having one or two integrated displays.

Binocular HMDs which have only one display typically show two views. The views are separated, whereby each eye is focusing on a different view, which enables monocular HMDs to offer the same stereoscopy effect as binocular HMDs with two displays. A comprehensive review of HMDs is given by Cakmakci and Rolland (2006). Since the release of the Oculus Rift, a low-cost HMD currently offering low-latency headtracking and a 75 Hz display with a resolution of 960 x 1080 pixels per eye (Development Kit 2), VR devices have received a lot of attention. The release followed announcements of big electronic companies, like Samsung and Sony, to release their own VR HMDs (Samung's Gear VR and Sony's Morpheus). In our study, a Oculus Rift is utilized in our system to track the user's head position and to render the graphical representation of our virtual scene.

2.2 Audio in Virtual Reality

Although the visual stimulus is probably the main point of interest of publications related to VR, there are a few publications describing the audio processing in more detail or propose new approaches for enhancing the listening experience of users (Astheimer, 1993).

The CAVE at RWTH Aachen University uses an audio rendering system based on binaural synthesis that uses loudspeakers for reproduction instead of headphones (Kuhlen et al., 2007; Schröder et al., 2010). When using a pair of loudspeakers for reproduction of binaural signals, it is intended that the audio signal of the left speaker is emitted only to the left ear and the audio signal of the right speaker is emitted only to the right ear. When reproducing binaural signals by loudspeakers, there is crosstalk between the loudspeakers, meaning that, e. g., audio signals from the left loudspeaker arriving also at the right ear. To overcome this problem they use dynamic crosstalk cancelation which suppresses the crosstalk between the loudspeakers.

In order to process an auditory representation of the virtual scene in real-time, their system uses a *fast convolution* which is a technique that also our VR system uses. The term fast convolution describes techniques that use a fast fourier transform (FFT) to convolve audio signals with BRIRs in the frequency domain (Stockham, 1966; Torger and Farina, 2001). Thereby, the convolution is performed by a multiplication of the discrete fourier spectra. Moreover, the convolution is processed block-wise, where each block contains a number of samples that correspond to the block length of the sound interface. Thus, the input-to-output latency is dependent on the block length, e.g., when having a sample rate of 48000 Hz, a block length of 4096 samples leads to a delay of

$\frac{4096}{48000\,\mathrm{Hz}} = 0.085\overline{3}$ seconds. A uniformly partitioned convolution splits BRIRs into multiple parts, where each part typically has the same number of samples equal to the block length. Besides the uniformly partitioned convolution, non-uniformly partitioned convolution also exists, where BRIRs are split into multiple parts with different numbers of samples (Gardner, 1994). By using parts with a number of samples higher than the block length, the computational effort is reduced compared to a uniformly partitioned convolution (Garcia, 2002). The system of Schröder et al. (2010) utilizes a non-uniformly partitioned convolution since their use-cases require simultaneously rendering a high number of virtual sound sources. Since our use-case requires only a limited number of virtual sound sources to be rendered, a uniformly partitioned convolution is used in our system.

DeFanti et al. (2009) argue that binaural approaches, especially headphone-based and head-tracked systems, are very useful for single-user scenarios but they are not well suited for multiple simultaneous users who may also want to converse with each other. The audio rendering of their system (StarCAVE) is achieved using surround speakers and wave field synthesis (WFS). The basic idea behind the WFS is based on the Huygens-Fresnel principle, which states that any wavefront can be assembled by a superposition of elementary spherical waves (Berkhout et al., 1993). In practice, this is achieved by a large array of independently controlled loudspeakers that is used to create the same pressure wave of a virtual sound source as an actual sound source located somewhere inside the sound field.

More work on the connection between audio and VR has been done in the field of auditory virtual environments (AVEs). AVEs are defined as the auditory components of virtual environments which aim at creating situations in which humans have perceptions that do not correspond to their physical environment but to the virtual one (Novo, 2005). A comprehensive overview of AVEs is given by Gilkey and Anderson (2014). Silzle et al. (2004) addressed the basic concepts of AVEs and described a comprehensive system with the purpose to generate AVEs. How a strong sense of presence is achieved by AVEs has been studied by Västfjäll (2003) and Larsson et al. (2004).

2.3 Real-world Experiments vs. Virtual Reality Experiments

Nowadays, VR systems are used in various fields, e. g., phobia therapy, military training, entertainment, and scientific experimental research (Loomis et al., 1999; Sanchez-Vives and Slater, 2005; Bowman and McMahan, 2007). In some of these works, VR experiments were carried out and their results were compared to a real-world experiment.

Bella (2004) conducted a social experiment to investigate the speed of vehicles while driving in work zone areas and compared the results of the VR experiment with the real-world scenario. The VR system he built was a driving simulator, aiming to be as similar as possible to an actual car. User interfaces (pedals, steering wheels, and gear lever) were installed on a real vehicle. The scenario of the construction site was projected onto three big screens: one in the center in front of the vehicle and two lateral ones angled at 60 ° with respect to the plane of the central screen. The whole setup was connected to a sound system to reproduce the sounds of the engine. The visuals and audio were rendered according to the traveling conditions of the vehicle, depending on the actions of the driver on the pedals and the steering wheel. The vehicles' speeds in the real-world work zone were measured by a laser speedmeter and compared to speed measurements obtained by the VR experiment. Bella's results show that samples from the VR experiment and the real-world scenario belong to the same population, i.e., no significant differences existed between the two environments in his scenario.

Gurusamy et al. (2008) investigated the effectiveness of VR trainings for laparoscopic surgery compared to other training methods including conventional (real-world) training. To this end, they compiled clinical trials that address laparoscopic surgery and analyzed the results. They came to the conclusion that VR trainings are helpful, especially for young surgeons at the beginning of their laparoscopic training, e. g., VR training reduced the operating time, error, and unnecessary movements during laparoscopic cholecystectomy (removal of the gallbladder). Moreover, there is convincing evidence that VR training is a useful supplement to conventional training in laparoscopic cholecystectomy for surgical residents with limited laparoscopic experience.

Vora et al. (2002) measured the degree of immersion and presence felt by subjects when conducting an aircraft visual inspection training in a VR simulator compared to the conventional PC-based training application. Although, PC-based training is not performed in the real-world, the work of Vora et al. shows how VR can be utilized to substitute conventional training methods. Their VR system was based on a HMD with six degree-of-freedom headtracking and used to create a virtual cargo bay environment. The results of their study show that the VR system scored well in most aspects of presence and was favored over the PC-based training.

More work on training has been done by Kozak et al. (1993), Witmer et al. (1996), Sveistrup et al. (2003), Rose et al. (2000), Bossard et al. (2008), and Psotka (1995).

As one can see, a lot work has been published to quantify the effect of VR in comparison to real-world tasks. To our knowledge, none of this work has mainly focused on auditory experiments. This paper is a contribution towards quantifying the effect of VR environments on auditory experiments.

2.4 Listening Room and Overall Listening Experience

The OLE (or a related attribute) is included in many models that aim to describe the process from starting with the sensation of an auditory stimulus and ending with a qualitative rating of the perceived listening experience. In 1972, Prince (1972) published a paradigm for describing this process. He presented his paradigm as a dependency graph showing which factors might be involved in a response to music and how they might be connected. His conclusion is that a wide range of factors (e.g., personality, maturity, musical ability, expectation, and even muscle movement) might be involved in the process of formulating a response and should be considered for research. Another model that was published by Blauert and Jekosch (2012) structure the formation process of *sound-quality* judgments. Their model is divided into four different layers: Auditive Quality, Aural-scene Quality, Acoustic Quality, and Aural-communication Quality. If one assigns the attribute OLE to their model, he or she would very likely assign it to the Aural-communication Quality layer. The Aural-communication Quality layer is described as the most abstract layer and its assigned attributes cannot be easily described by physical measures or features of an audio signal. Schoeffler and Herre (2014a) proposed a model that allows various other models to be integrated and can be used to predict OLE ratings. Their model includes the sensation, perception, and cognition of an auditory stimulus and the personality, ability, and state of a listener.

The OLE, being a very holistic attribute, is considered to be influenced by many factors, but only a few studies were conducted to investigate these factors of influence on the OLE. For example, degradations in audio quality (like distortions in frequency domain or bandwidth-degradation) have a strong effect on the OLE (Schoeffler and Herre, 2013; Schoeffler et al., 2013a; Rumsey et al., 2005). Furthermore, the individual listener, including his or her personality, has been shown to have a significant influence on the OLE (Pearson and Dollinger, 2004; Schoeffler and Herre, 2014b).

Related to spatial audio, the influence on up- and down-mix algorithms has been investigated by Schoeffler et al. (2014a). An up- or down-mix algorithm is needed when an audio source material has fewer, or more respectively, channels than the reproduction system. For example, a down-mix algorithm is needed when 5.1 surround material is played back by a stereo reproduction system. If no down-mix algorithm is applied, information contained in the center and surround loudspeakers would be discarded. In their study, participants rated the overall listening experience while listening to up- and down-mixed music. The results of the study indicated that the down- or up-mix algorithm has only a minor influence on OLE ratings. There was one exception, a low-quality up-mix and a low-quality down-mix that were considered to be lower anchors in the study. In another study, Rumsey et al. (2005) showed that surround sound is very important for preference ratings given by naïve listeners. Furthermore, they proposed a regression model fitted with their experiment results that predicts preference ratings of naïve listeners based on the timbral quality and spatial quality ratings of expert listeners. The findings of Rumsey et al. were confirmed by an experiment of Schoeffler et al. (2014b) where the influence of single-/multi-channel systems (mono, stereo, and 5.1 surround sound) used for reproduction was subject to investigation. In their study, participants rated the overall listening experience while listening to music reproduced by different reproduction systems. The mono, stereo, and 5.1 surround sound system had significant influences on the OLE ratings. In particular, the mono system had the weakest effect and the 5.1 surround sound system had the strongest effect. In the same study, the effect of the listening room was investigated. The study consisted of two main experiment sessions. In the first experiment session, listeners were asked to rate the OLE while sitting in a professional listening room and listening to a short music excerpt reproduced by mono, stereo, and 5.1 surround sound. Two and a half months later, the same experiment was conducted but this time listeners sat in a common office room (second experiment session). In both sessions, participants were sitting inside a black-colored 360 ° masking curtain made of deco-molton that was installed to veil the loudspeakers and the appearance of the room. By using a masking curtain, the experimenters controlled the influence of the visual stimulus on the OLE and expected that the participants would focus on the acoustic characteristics of the room. The office room had a reverberation time that would have

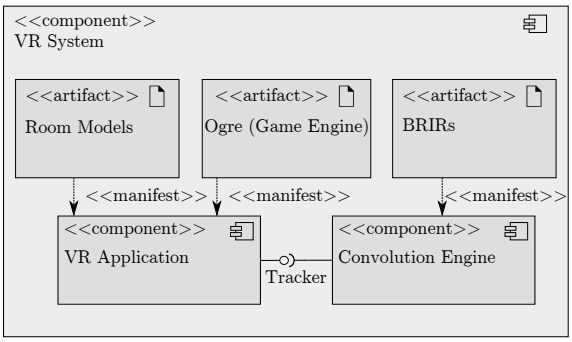

Fig. 1: Software architecture of the VR system depicted as an UML component diagram.

been rated by acousticians very low compared to the professional listening room. However, comparing the results of both sessions, no significant differences were found. Although the experiment presented in this paper focuses on investigating the effect of VR environments on experiments, the experiment also continues the study of Schoeffler et al. by investigating the differences between two rooms without veiling the rooms' visual appearance (Hypothesis II).

3 System for VR Experiments

3.1 Overview

Our VR system consists of two main components: the VR Application and the Convolution Engine (see Figure 1). The main purpose of the VR application is to process the user's input and to render the visual representation of the virtual scene on the display of the HMD. Thus, the VR application contains the whole business logic, e. g. showing the instructions to the participants, controlling the workflow of the experiment, and fetching sensor data from the HMD. The second main component is the Convolution Engine which retrieves head-tracking data from the VR Application and renders the auditory stimuli. The system we present does not contain any major new algorithms that we want to propose to the VR community. The components use state-of-the-art algorithms or common practices interacting with each other in order to create an authentic VR environment.

3.2 VR Application

The VR Application must render virtual scenes in which a participant of an experiment is sitting inside a room and taking part in an auditory experiment. In order to authentically render appearance of a room,

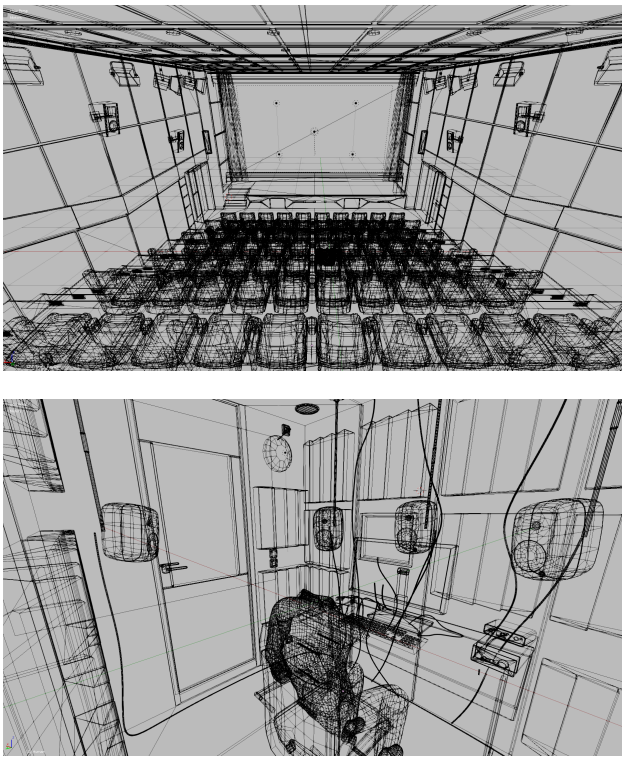

Fig. 2: Wire-frame view of the cinema model (upper image) and listening booth model (lower image).

we created models of the two rooms we used in the experiment. The first room was a medium-sized cinema and the second one was a small listening booth, both located at the venue of the Fraunhofer IIS in Erlangen, Germany. More details about the rooms are given in Section 4.1.3. Both models were modeled true-to-scale, since we were using blueprints of the rooms. Figure 2 depicts wire frame views of the two models.

For rendering the models, we used a game engine called OGRE (The OGRE Team, 2013). The main reason for using OGRE was that it is completely open-source. When we started to implement our VR system, no graphics engine had official support for the Oculus Rift. Therefore, we relied on modifying parts of the source code of a graphics engine in order to render a virtual scene into two views (of a HMD). In addition, OGRE has also an active community, which made it the optimal choice for us at that point in time.

Besides rendering the room appearance, the VR Application enables participants to read instructions and to listen to auditory stimuli and rate them. In typical real-world experiments, a graphical user interface (GUI) (or sometimes sheets of papers) is used where participants can read instructions and rate stimuli. To have the same experience in VR, we programmed a GUI framework (based on Gui3D (Frechaud, 2013)) that al-

lowed us to create a virtual screen within the virtual environment. On this screen, we could place GUI elements (buttons, textfields, audio players, etc.) into an interface similar to the one used in the real-world experiment. For our experiment, we placed such a virtual screen at a monitor model of the VR listening booth. In the VR cinema, we placed the virtual screen at the cinema screen.

One could design the same GUI for a VR experiment as would be used in a real-world experiment. However, due to technical limitations of the Oculus Rift (and other VR HMDs) having the same look&feel can lead to unexpected results. Nowadays, the resolution of a typical monitor is about 1920 × 1080 pixels. Assuming the GUI of the real-world experiment is shown in full-screen, all pixels are used for presenting the GUI. The display of the current version of the Oculus Rift has a resolution of 1920 × 1080 pixels. As the display is split into two views (for each eye one view), the maximum resolution for rendering the virtual scene is limited to 960 × 1080 pixels. Moreover, only a small area of the provided resolution is used for rendering the virtual screen that shows the GUI. In the VR listening booth, the actual resolution that is used for rendering the virtual screen is about 550 × 325 pixels, considering the participant is sitting in front of the monitor and looking towards the monitor. Sitting in the VR cinema, about 500 × 295 pixels are used for rendering the virtual screen. In addition to having only a very limited number of pixels to render a GUI in the VR environment, the so-called screen-door effect (or fixed-pattern noise) deteriorates the appearance of the GUI. The screen-door effect is a visual artifact in which the fine lines separating the display's pixels become visible in the rendered image, especially on white backgrounds. Such an effect is very present on HMD since users' eyes are very close to the display. Moreover, HMDs have in most cases a low resolution which enhances this effect. When a GUI of 1920 × 1080 pixels is scaled to 550 × 325 or 500 × 295 pixels and the screen-door effect is present, the GUI significantly loses the intended look. To overcome this problem, one could design a GUI having only very simple control elements that use a lot of space. We tested such an approach in a usability test, where very basic GUIs were shown on a display in a real-world environment. The participants reported that it felt very "unnatural" to use such a GUI, e. g., where a button has the width and the height of about 10 % of the total display resolution. Therefore, we decided to design two different look&feels for the experiment, where great care was taken to present the same information by the two different GUIs. The differences are

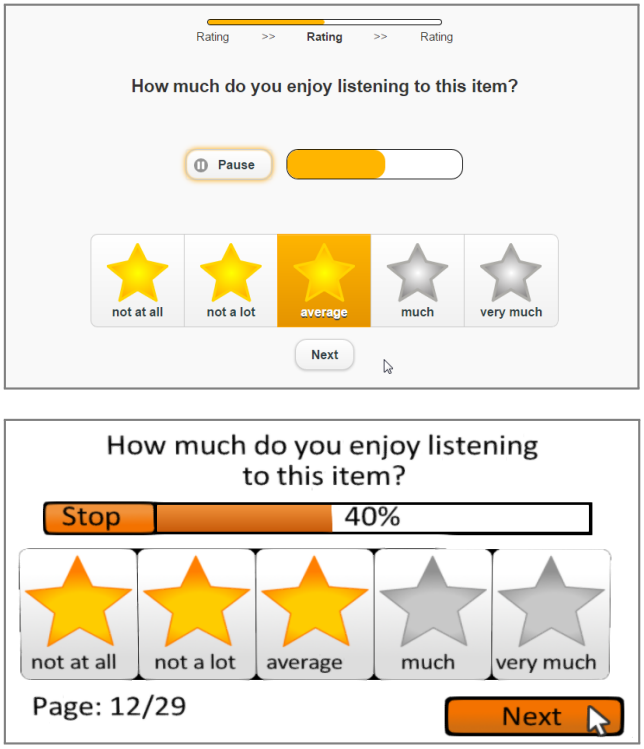

Fig. 3: Screenshots taken from the real-word experiment software (upper image) and the VR experiment software (lower image).

depicted in Figure 3, where screenshots of the two different GUIs are shown.

In our system the mouse is used to give responses by clicking on control elements shown by a screen of a virtual display. Another feature of our VR application is that mouse-movements of the user in the real-world are also shown in the VR environment. The reason for integrating this feature is that many participants of a pilot experiment reported that the visual and auditory representation was very satisfying but it just felt unnatural to move the mouse and not to have any visual feedback in the VR environment. In order to implement a visual feedback, we used the x- and y-coordinates of the mouse pointer which could be retrieved from the operating system. The values of the coordinates were converted to a mm-scale of the real-world. For example, if the value of the x-coordinate was about the width of the screen, the virtual mouse was moved to the right-end of the area were the virtual mouse was located. This basic approach has some limitations, so, e. g., if someone lifted or moved the mouse very far away, the location of the visual mouse would not correspond to the location of the real-world mouse. A more advanced approach would require tracking the coordinates of the real-world by a camera or sensor located inside the mouse, but results of

another usability test indicated that the used approach would be sufficient for the final experiment.

3.3 Audio Rendering

3.3.1 BRIRs Measurements

In general, BRIRs are measured by placing a dummy head with two in-ear microphones at the listening position and playing back a stimulus (excitation signal) from a loudspeaker. The stimulus is recorded by the two microphones resulting in two signals from which the BRIRs are extracted. There are several approaches to do this, where each approach has its advantages and disadvantages. Nowadays, it is common practice to use a variant of the logarithmic-sweep method, since these kinds of methods provide excellent signal-to-noise ratios. The BRIRs that are used for rendering the auditory stimulus in the VR rooms were measured with a logarithmic-sweep method based on the approach proposed by Müller and Massarani (2001).

For the VR environment, we measured BRIRs with a dummy head in both rooms at the same location where the virtual avatar of the participant would sit. In order to achieve satisfying immersion, we had to measure BRIRs at different head rotations. The reason for this is that the localization of sound sources (or spatial hearing in general) is strongly influenced by perceived differences between the ears. For example, the localization of sound sources is, among other factors, dependent on the so-called interaural time difference (ITD) and interaural level difference (ILD) (Palomäki et al., 2005; Sandel et al., 1955). The ITD is the difference in arrival time between two ears of two sounds. If a sound arrives earlier at the right ear, we instinctively assume that the sound is coming from the right. The ILD, sometimes also called interaural intensity difference (IID), is similar to the ITD but describes the differences in loudness between the two ears. If a sound source is located at the front of the listener and the listener moves his or her head, the ITD and ILD change.

Since participants usually move their heads while sitting in an auditory experiment with loudspeakers, head movement must be supported by a VR system. Moreover, sound sources should be perceived as similar as possible to the way they are perceived in the real-world while moving. Therefore, we measured BRIRs at different head positions with a custom-made dummy head (designed by Hess and Weishäupl (2014)) that supports neck and head movements in the three rotational degrees of freedom. The dummy head has a mechanical compression spring which allows a range of motion of ±30° for pitch and roll. Four stepper motors are used to tilt a base board holding the dummy head in which two microphones (DPA 4061) in the ear-canals are incorporated for binaural measurements. This construction is placed on a rotatory actuator, a combination of stepper motor and turntable. For rotation in the horizontal plane a full turn would be possible, but it is limited by software to ±90°. The dummy head consists of many parts, including motors and microphones, that influence the measured impulse responses. Great care was therefore taken to obtain impulse responses as authentic as possible (e. g. by powering off the motors during the sweep playback).

The dummy head was programmed to automatically move its head according to a list of configured head rotations. Moreover, the head was connected to the sound system of the room to be measured, allowing automatically triggering measurement at each position. The dummy head was configured to measure BRIRs with a length of 32768 samples using a sample rate of 48000 Hz. The resulting length of each impulse response was $\frac{32768}{48000\,\mathrm{Hz}} = 0.683$ seconds. The range and resolution of head rotations measured were different for yaw, pitch, and roll. The dummy head yawed from −40° to 40° with a step size of 1°. Pitch movement ranged from −6° to 6° and roll movement ranged from −3° to 3°. Both movements used a step size of 3°. The reason for introducing a wider range in yaw than in pitch and roll was that changes in yaw have a stronger influence on human spatial perception (Blauert, 1997), especially if a surround sound setup without height loudspeakers is used. The same applies to the step sizes used. The step size in yaw was chosen to be smallest due to its importance for the spatial perception. The maximum step sizes of 3° were chosen as they lead to adequate results and are a very good compromise regarding the time needed for measuring all BRIRs (Lindau et al., 2008). We recorded the 5.1 surround sound system that was installed in each room, resulting in six input channels.

The user experience would have been significantly improved if individual/person-specific BRIRs had been used (Väljamäe et al., 2004). When individual BRIRs are measured, the person itself, or an accurate replica of his or her ears, head, and torso, must be used during the measurement. Due to the high number of participants and the high number of sessions, using individual BRIRs was considered to be too time-consuming. Moreover, in total, 1215 different head rotations and six different channels were measured which took about ten hours for one room. Therefore, if individual BRIRs had been measured, a person would have had to spend the same amount of time plus additional time for orienting the person's head, since each measurement had to be accurate to much less than one degree.

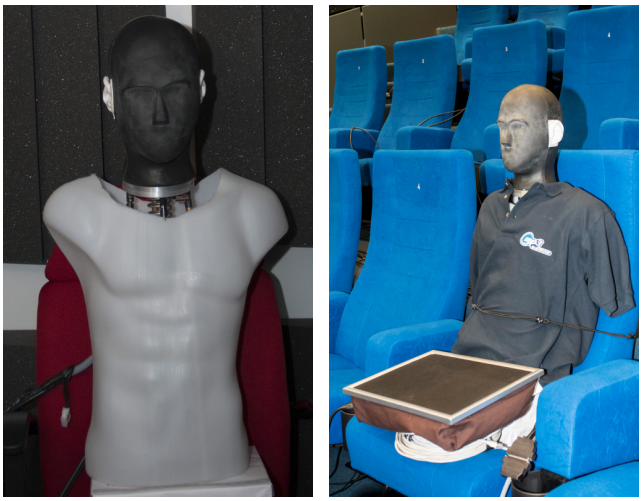

Fig. 4: The dummy head used for the BRIRs measurement. On the left hand side, the torso and a part of the four stepper motors can be seen. The right hand side shows the dummy head while conducting a measurement. The lap mouse pad in front of the dummy head was also used in the real-world sessions.

3.3.2 Convolution Engine

We implemented a convolution engine that allowed us to convolve any number of input signals and BRIRs (Schoeffler and Hess, 2012). As a 5.1 surround sound format was used in the experiment, we configured the convolution engine to convolve six input channels with six BRIRs (filters) for each ear. The convolution engine used is based on a fast convolution, so multiplication in frequency domain is applied (a detailed introduction into convolutions in frequency domain is given by Lathi and Green (2014)). The following equations show how our convolution engine works in detail.

The convolution engine is parameterized by the following attributes:

- s ∈ {left, right} is the side of the headphone
- p ∈ {left, right, center, ...} is the loudspeaker channel
- block number n
- r_n is the position and rotation (in all three axes) of the listener's head at block number n

The output signal y^{s} is dependent on the input signal x^{p} of indefinite length and the impulse response $h^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}$ of finite length L_H. Since the signals are block-wise processed, the input signal x^{p} is divided into blocks, where each block consists of a number of samples equal to the block length of the audio interface.

The input signal consists of blocks x_i^{p} with block length L_I

$$x^{\mathrm{p}} = \left[x_0^{\mathrm{p}}, x_1^{\mathrm{p}}, x_2^{\mathrm{p}}, \ldots\right], \tag{1}$$

where i is the block index. Next, the filter $h^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}$ is divided into N sub filters $h_j^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}$ ($j \in \{0, 1, \ldots, N-1\}$) with length L_I

$$h^{\mathrm{p},\mathrm{r}_n,\mathrm{s}} = \left[h_0^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}, h_1^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}, h_2^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}, \ldots, h_{N-1}^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}\right]. \tag{2}$$

L_I does not need to be a factor of L_H since in case of $N \cdot L_I > L_H$ zero-padding is applied

$$h_{N-1}^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}[u] = \begin{cases} h^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}[w+u] & \text{if } w+u < L_H \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

where $w = (N-1) \cdot L_I$ and $u \in \{0, 1, \ldots, L_I - 1\}$.

To prevent aliasing due to circular convolution, $h_j^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}$ is zero-padded to length $2L_I$

$$h'^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}_j = \left[h_j^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}, 0, 0, \ldots\right]. \tag{4}$$

The signal block x'^{p}_i is expanded to length $2L_I$

$$x'^{\mathrm{p}}_i = \begin{cases} [0, 0, \ldots, 0, x_i^{\mathrm{p}}] & \text{if } i = 0 \\ [x_{i-1}^{\mathrm{p}}, x_i^{\mathrm{p}}] & \text{otherwise} \end{cases}, \tag{5}$$

where $i \in \{0, 1, 2, \ldots\}$.
Next, the filter and the input signal are transformed into frequency domain

$$H_j^{\mathrm{p},\mathrm{r}_n,\mathrm{s}} = \mathrm{FFT}\left\{h'^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}_j\right\} \text{ and } X_i^{\mathrm{p}} = \mathrm{FFT}\left\{x'^{\mathrm{p}}_i\right\}. \tag{6}$$

Since the filtering is linear, successive blocks can be filtered with the corresponding sub-filter one at a time and the output blocks are fitted together to form the overall signal

$$y'^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}_n = \mathrm{IFFT}\left\{\sum_{k=0}^{N-1} X_{n-k}^{\mathrm{p}} \cdot H_k^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}\right\}. \tag{7}$$

X_{n-k}^{p} is set to zero if $n - k < 0$.
$y'^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}_n$ has length $2L_I$. While the first half is the overlapping part of the convolution and is discarded, the second half contains the required output block of one channel p at one position r_n

$$y_n^{\mathrm{p},\mathrm{r}_n,\mathrm{s}} = \left[y'^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}_n[L_I], y'^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}_n[L_I+1], \ldots, y'^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}_n[2L_I-1]\right]. \tag{8}$$

Due to the fact that there are measurements of impulse responses for only a finite number of rotations r, the closest r to the actual rotation is chosen. If the value of the current r, r_n, is different from the previous r_{n-1},

the output signal y_n^{s} is computed by a squared cosine cross-fading of $y_n^{\mathrm{p},\mathrm{r}_{n-1},\mathrm{s}}$ and $y_n^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}$

$$y_{n,\mathrm{weighted}}^{\mathrm{p},\mathrm{r}_{n-1},\mathrm{s}}[u] = y_n^{\mathrm{p},\mathrm{r}_{n-1},\mathrm{s}}[u] \cos^2\left(\frac{\pi \cdot u}{2L_I}\right) \tag{9}$$

$$y_{n,\mathrm{weighted}}^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}[u] = y_n^{\mathrm{p},\mathrm{r}_n,\mathrm{s}}[u] \left(1 - \cos^2\left(\frac{\pi \cdot u}{2L_I}\right)\right) \tag{10}$$

where $u \in \{0, 1, \ldots, L_I - 1\}$.
Finally, the output signal is calculated

$$y_n^{\mathrm{s}} = \begin{cases} \frac{1}{P}\sum_{\mathrm{p}} y_n^{\mathrm{p},\mathrm{r}_n,\mathrm{s}} & \text{if } \mathrm{r}_{n-1} = \mathrm{r}_n \\ \frac{1}{P}\sum_{\mathrm{p}} y_{n,\mathrm{weighted}}^{\mathrm{p},\mathrm{r}_{n-1},\mathrm{s}} + y_{n,\mathrm{weighted}}^{\mathrm{p},\mathrm{r}_n,\mathrm{s}} & \text{otherwise} \end{cases}, \tag{11}$$

where P is the number of loudspeaker channels.

4 Experiment

4.1 Method

The purpose of the experiment was to test Hypothesis I ("The OLE ratings given in the VR environment do not differ from those given in the real-world experiment") and Hypothesis II ("The OLE ratings that were given while being present in the (real-world) cinema do not differ from those given in the (real-world) listening booth."). To this end, participants joined five sessions, where they rated the OLE of music excerpts. In the first session (basic item session) they rated the OLE of music stimuli in a "neutral room" with headphones. These ratings were expected to reflect how much a participant likes a specific stimulus without being influenced by the room or the environment. In the other four sessions, participants sat in the cinema or listening booth either in the VR or in the real-world environment.

4.1.1 Participants

Thirty participants (24 males, 6 females) volunteered to participate in the experiment. Twenty participants reported an age between 20 and 29 years, nine reported being between 30 and 39 years old and one participant reported being between 40 and 59 years old. Twenty participants identified themselves as professionals in audio (audio researchers, audio engineers, Tonmeisters, etc.). Twenty-seven participants were familiar with listening tests and indicated that they had volunteered for at least one listening test before. Twelve participants reported that they regularly play computer games for more than one hour a week.

Song	Interpreter
You're Beautiful	James Blunt
She Drives Me Crazy	Fine Young Cannibals
Everyday Is A Winding Road	Sheryl Crow
Cold As Ice	Foreigner
Messias	Georg Friedrich Händel
In My Head	Jason Derulo
Ironic	Alanis Morissette
Symphony Nr. 4	Peter I. Tschaikowsky
Have You Ever Seen The Rain	Creedence Clearwater Revival
Long Train Runnin'	The Doobie Brothers
Amazing	Seal
Shout	Tears For Fears
Chase The Thrill	Nikka Costa
Tonight	Alex Max Band

Table 1: Selected music excerpts of the experiment.

4.1.2 Stimuli

Two sets of stimuli were used in the experiment. The first set was used for giving some indication of how much the responses of listening-room- and reproduction-system-dependent attributes differ between the VR and real-world environment. The first set contained the following stimuli:

pink noise The pink noise signal (peak = −10.4 dB, crest factor = 12.3 dB) had a duration of ten seconds and was rendered in mono. It was thus played back by the center channel of the surround systems.

castanets The castanets signal was recorded dry and had a duration of about seven seconds. The castanets recording was mixed in stereo. It was thus played back by the left and the right channel.

drums The drums recording was a seven-second-long beat where mainly the bass drum, toms, and cymbals were played. The drums recording was also mixed in stereo.

The second set of stimuli was used to rate the OLE and contained fifteen music excerpts of songs of various genres (see Table 1). The songs were obtained from the "Mercedes-Benz Signature Sound" DVD and the "BR Klangdimensionen" DVD. The excerpts had a duration of about ten seconds and mainly covered the most recognizable part of the song (e. g., the refrain). The stimuli of the second set were originally mixed in 5.1 surround sound. In the first session, the basic item session, the participants used headphones, so stereo-downmixes of the stimuli were played back.

In order to have all stimuli within the same range of loudness, an EBU-R128 loudness-normalization was applied (European Broadcasting Union, 2011).

4.1.3 Materials and Apparatus

Software Infrastructure The undertaking of the experiment required a complex software infrastructure

since many components were involved and sessions had to be taken by participants in parallel. The main reason for parallel sessions was that in total 150 sessions (30 participants × 5 sessions) had to be supervised and the rooms could only be booked for a limited period of time. The software infrastructure mainly consisted of the VR Application, which was already described in Section 3 and a Web Application, which was used for the basic item session and the two real-world sessions. An overview of the complete software infrastructure as a UML component diagram is shown in Figure 5.

The Web Application is mainly a framework that was developed at the International Audio Laboratories Erlangen for the purpose of easily and time-efficiently creating GUIs for experiments by using web technologies. The Web Application was deployed in two variants: The first variant was used in the basic item session and the second variant was used in the two real-world sessions. The difference between the two variants was that the basic item session had a totally different procedure than the two real-world sessions.

All three applications accessed the same data source that contained the music excerpts that were used as stimuli in the experiment. The Web Application of the basic item sessions retrieved the stereo down-mixes as they were used during the first session. The other two applications fetched the original 5.1 surround sound mixes. The responses of the participants and all other data needed was stored into and retrieved from a MySQL database. Furthermore, the database component executed consistency checks, e. g., checking that the same session was not performed twice by a participant and confirming that participants took the sessions in the intended order.

Professional Listening Room The participants sat in a professional listening room during the basic item session and the VR sessions.

In the basic item session, participants sat at a desk, where a 24" widescreen LCD monitor, mouse, and keyboard were placed. The monitor displayed the experiment software. The participants used Beyerdynamics DT 770 PRO headphones connected to a LAKE People Phone AMP G109 amplifier.

The audio equipment used for the VR environment sessions is described in the next paragraph.

Virtual Environment The VR sessions were done on a high-end PC on which the VR Application was running. A PreSonus Audio Box 44VSL sound interface was connected to the PC. Participants used electrostatic headphones (Stax SR-507) driven by a Stax SRM 600 amplifier which was connected to the

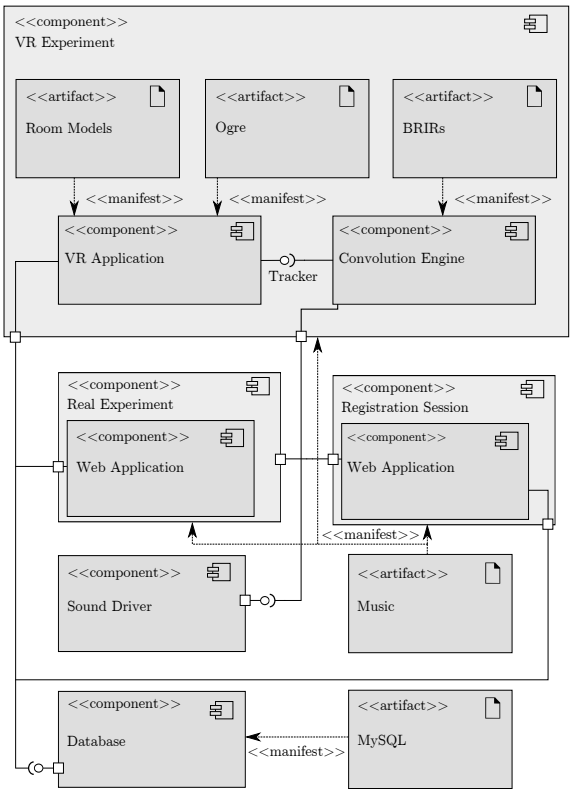

Fig. 5: Software architecture of all systems used in the experiment.

PreSonus sound interface. The measured BRIRs were post-processed to compensate the headphone's transfer function by inverse filtering.

As mentioned before, the Oculus Rift HMD was used to render the images of the virtual scene and to track head positions of the participants. The display of the Oculus Rift had a resolution of 960 × 1080 pixels per eye, and could be set to a refresh rate of 60, 72, and 75 Hz. For the virtual sessions, we set the refresh rate to 75 Hz. The orientational head-tracking is based on data from a gyroscope, accelerometer, and magnetometer and has an update rate of 1000 Hz.

The Oculus Rift also supports positional tracking which was not fully enabled in our VR application, since we measured only BRIRs with different orientational positions. The VR application allowed only translational movements of ±10 cm in normal direction of the median, frontal, and horizontal planes. If a participant moved his or her head more than ±10 cm in any of these axes, the virtual avatar of the participant stopped moving its head. In previous usability tests, stopping the movement was well recognized as a feedback that translational movements are not fully supported, and the limit of 10 cm turned out to be a good trade-off.

Fig. 6: Screenshots of the virtual scenes as they were displayed on the Oculus Rift. The upper screenshot shows the VR cinema and the lower screenshot shows the VR listening booth. Each screenshot shows two views of the scene: the left one rendered for the left eye and the right one rendered for the right eye.

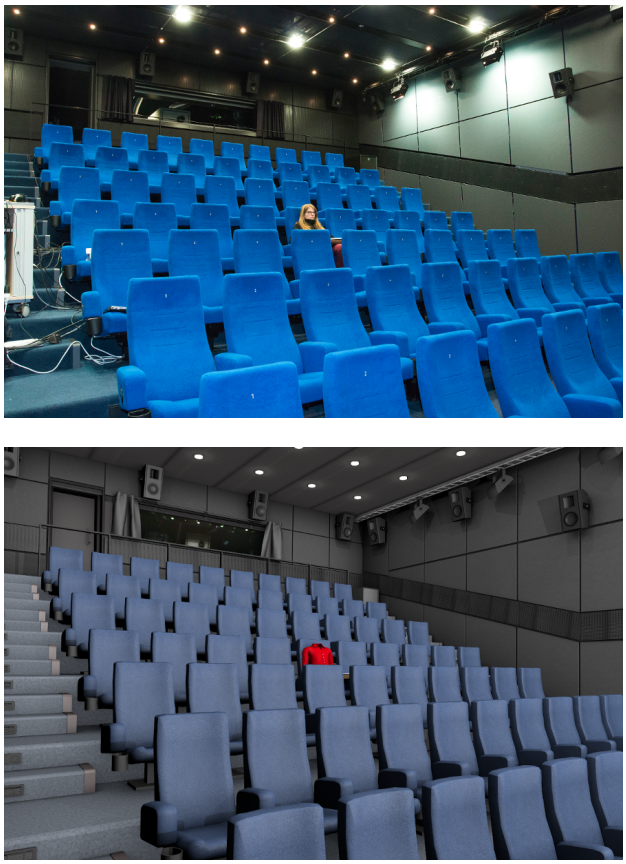

Fig. 7: The upper image shows the real-world cinema and the lower image shows the VR cinema.

The stereoscopic effect of the Oculus Rift can be optimized by configuring the interpupillary distance (IPD) of a person. Since it would have been too time-consuming to measure the individual IPD of each participant, we defined and used only two profiles (male and female). The male profile had an IPD of 64.5 mm and the female profile had an IPD of 62.5 mm.

Figure 8 and Figure 7 depict the VR listening booth and the VR cinema. Figure 6 depicts two screenshots that show how the virtual scenes of the cinema and listening booth were rendered on the display of the Oculus Rift.

Listening Booth The listening booth had room measurements of 1.81 × 2.44 × 2.07 m and provided enough space for one participant. A desk was placed inside the listening booth on which a monitor, keyboard, and mouse was placed. The monitor was used to display the experiment software. Inside the listening booth, a 5.1 surround sound system was installed with five Genelec 8030 APM speakers and one Genelec 7050B subwoofer. The speakers and the subwoofer were controlled by a SPL Surround Monitor Controller (Model 2489).

The reverberation times of the listening booth are shown in Figure 10. A picture of the listening booth is shown in Figure 8.

Cinema The cinema had room measurements of 12.3 × 8.75 × 5.6 m (height measured at stage) and a capacity of 70 seats (7 rows with 10 seats each). The cinema had installed multiple sound systems. We used the 5.1 surround sound system which is based on the Alcons CRMS (Cinema Ribbon Monitor) integrating a 3-way main channel (left/center/right) and a 2-way surround sound system. Since the system must provide an acoustical sweet spot wide enough for an entire audience, six loudspeakers are used for each surround sound channel. The aspect ratio of the cinema screen was set to 16:9 and room lighting was rather darkened. As in the VR cinema, every participant sat at the fifth seat in the fourth row. Since participants had to use a physical mouse during the experiment, a mouse pad was provided that could be laid on the lap.

The reverberation times of the cinema are shown in Figure 10. A picture of the cinema is shown in Figure 7.

Fig. 8: The left image shows a picture of the real-world listening booth. The right image shows the corresponding picture from the VR listening booth.

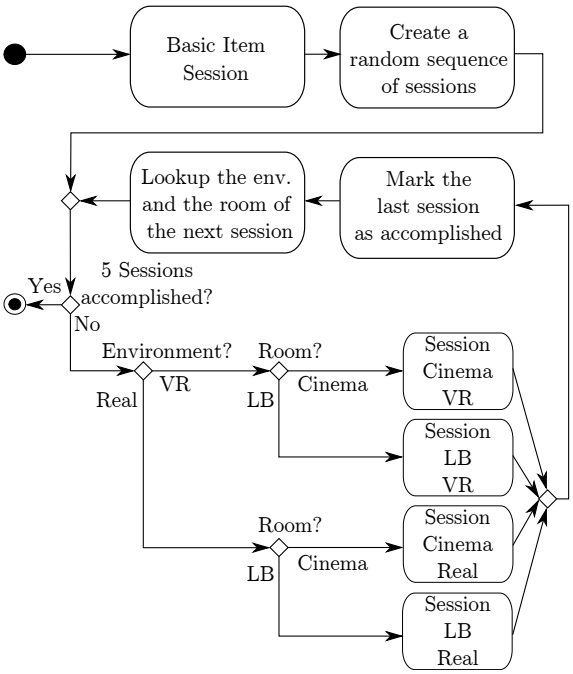

Fig. 9: The procedure of the experiment depicted as a UML activity diagram.

Loudness Equalization Since the loudness level of music being played back might have a significant impact on the listening experience, the loudness was equalized between the real-world environment and the VR environment. A pink noise stimulus (peak = −0.7 dB, crest factor = 12.8 dB) was used for all loudness measurements. The stimulus was recorded with a dummy head (Cortex Manikin MK1) which had one microphone installed in each ear.

In the first session, the participants listened to stereo stimuli with headphones. The loudness of the playback system was calibrated to 80 dBA SPL for each microphone.

The loudness of the virtual environment and real-world environment was measured separately for each channel. The left and surround left channels were measured by the microphone installed in the left ear of the dummy head. The other channels (right, center, LFE, and surround right) were measured by the right ear of the dummy head. The loudness of each channel was calibrated to 70 dBA SPL, except the LFE channel which was calibrated to 50 dBA SPL. The main reason for calibrating the LFE channel to a lower loudness was that the LFE channel plays back only the lower frequency range of a signal. High volume in such a low frequency range leads to audible artifacts when reproduced by headphones.

4.1.4 Procedure

The participants had to accomplish five sessions in total. An overview of the procedure is depicted as a UML activity diagram in Figure 9.

The first session, the basic item session, took place in the professional listening room. Participants sat in front of a computer that ran the experiment software. All instructions were presented by the experiment software and participants responded only by using the experiment software. In the basic item session, headphones were used for listening to the stimuli. At the beginning of the session, participants filled out a questionnaire. They were asked their gender, their age group, whether they are professionals in audio (e. g., sound engineers, audio researchers), and about their familiarization with listening tests. In addition we asked them whether they play computer games more than one hour a week and whether they listen via headphones for more than one hour a week.

The reason for asking the latter two questions was that the whole procedure of the VR sessions probably slightly resembles computer games. Therefore, participants who often play computer games might find it easier to interact with the VR environment. The last question about the headphones was asked because one major difference between the real-world and VR environment is that participants wore headphones in the VR environment. In case headphones would have de-

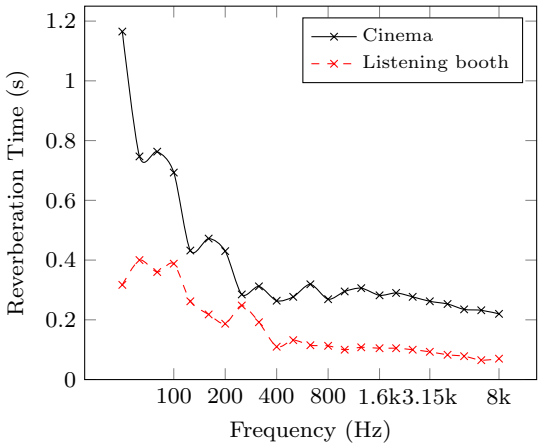

Fig. 10: Reverberation time (RT 60) of the listening booth and the cinema.

creased the degree of immersion, we wanted to check whether this is especially true for participants that are not used to wearing headphones.

After answering all questions, participants read the instructions about rating the OLE of music excerpts subsequently presented. The instructions stated that participants are asked to rate each music excerpt according to how much they enjoy listening to it. In addition, it was emphasized that we were interested in participants' personal opinion and that they should take everything into account in their rating that they would take into account in a real-world (non-experiment) scenario. The reason for giving the participants this additional instruction is that the majority of listening tests conducted at our institute are about assessing audio quality. In order to avoid that a participant accidentally rates the audio quality of the items, we added this additional instruction.

After reading the instructions, participants were asked to rate the OLE of fourteen music excerpts in a multi-stimulus comparison, which means that all music excerpts were presented on the same screen. As mentioned in Section 4.1.2, the stereo down-mixes of the stimuli were played back in this session. The order in which the music excerpts were presented was randomized. The question that was asked to the participants was "How much do you enjoy listening to each item?". Participants responded by using a five-star Likert scale which was labeled with "not at all", "not a lot", "average", "much", and "very much". Each music excerpt could be played back as often as desired. In order to rate a music excerpt, participants had to completely play back the music excerpt at least once. In the remainder of the paper, the ratings retrieved from the first session are called *basic item ratings*.

Next, the participants filled out another questionnaire where feedback could be given to the experimenters. At the end of the first session, each participant had to spend a few minutes trying out a demo application of the Oculus Rift. The purpose of playing around with the Oculus Rift was to familiarize the participants to better avoid cybersickness during the VR sessions. Furthermore, we tested whether each participant could clearly see using either the male or female IPD profile.

The basic item session was followed by another four sessions: one session in the real-world cinema, one in the real-world listening booth, one in the VR cinema, and one in the VR listening booth. Between each session participants had to take a break of at least a few hours. If possible, the next session was taken one day after the last session. The reason for having these long breaks was that participants might get annoyed or bored when they listen to the same song several times over a short period of time. The four sessions had almost the same procedure. One major difference between the real-world sessions and the VR sessions was that the GUI of the experiment software had a different look&feel, which was already mentioned in Section 3.2. The real-world cinema and real-world listening booth sessions took place in the respective rooms. The VR sessions took place in the professional listening room which was also used for the first session.

In all four sessions, the experimenter was not in the room, so all instructions were given by the experiment software and participants gave all responses using the experiment software. The experimenters were only present at the beginning of the VR sessions to help the participants setting up the Oculus Rift and to make sure that they wore the headphones with the correct orientation.

Each session was divided into two parts. In the first part, questions were asked that were related to room acoustics. In the second part, participants rated the OLE of music excerpts.

At the beginning of each of the four sessions, instructions were shown to the participants. The instructions gave some information that the following questions are related to the room acoustics and the perception of sound. The first question asked was how loud they perceive the presented stimulus (pink noise). As with the other stimuli, participants were allowed to play back the stimulus as often as desired. To report the loudness, participants used a Likert scale with the values "very quiet", "quiet", "normal", "loud", and "very loud". Next, the stimulus "castanets" was presented and participants were asked how much reverb the room had. The question about reverb was answered on a Likert scale with the values "none", "a little", "average",

"much", and "very much". Subsequently, the participants had to indicate how far away they perceive a stimulus (pink noise). They reported the location by a Likert scale with the values "very near", "near", "average", "far", and "very far". Then, two questions followed where participants were asked how much they like the bass and the treble of a stimulus (drums). The participants gave both ratings on a Likert scale with the values "not at all", "not al lot", "average", "much", and "very much". When the last question was answered, the second part of the session started.

Again, instructions were shown to the participants which were very similar to the instructions shown in the basic item session. They were instructed to rate the OLE of short music excerpts and it was again emphasized that we were interested in participants' personal opinion. Next, the fourteen music excerpts were presented in a single-stimulus comparison, i. e., each music excerpt was separately rated. The question we asked to the participants was "How much do you enjoy listening to this item?". Participants gave their responses on the same five-star Likert scale used in the first session. In order to rate a music excerpt, participants had to completely play back the music excerpt at least once. A screenshot of the GUI as it was used in the real-world sessions and in the VR sessions is shown in Figure 3. In case the session was a VR session, participants were asked whether they felt dizzy while doing the session. This question could be answered with "no", "a bit", or "yes".

4.2 Results

The most important dependent variable, item rating, is strictly speaking an ordinal variable. Therefore, the hypotheses are mainly verified by non-parametric statistics. Nevertheless, due to predominant use of parametric statistics and since the item ratings can also be interpreted as an interval or ratio variable (number of stars), parametric statistics might additionally be used to confirm the results of the non-parametric statistics. Moreover, since significance levels do not provide enough information about the practical or theoretical importance of an effect, effect sizes are also reported (Fritz et al., 2012). Throughout the paper, Pearson's r is used as an effect size (Fritz et al., 2012) when non-parametric statistics (e. g. Wilcoxon signed-rank test) were applied. The values of r can vary from -1 to 1, -1 indicating a perfect negative relation, 1 indicating a perfect positive relation, and 0 indicating no relation between two variables. The strength of an effect is interpreted according to Cohen's guidelines (Cohen, 1988), with $r = 0.1$ is a weak effect, $r = 0.3$ is a moderate effect and $r = 0.5$ is a strong effect. In the case that parametric statistics (e. g. paired t-test) were applied, Cohen's d is used as an effect size metric. The meaning of the effect size value is interpreted according to the standard interpretation: $d = 0.2$ is a small effect, $d = 0.5$ is a moderate effect, and $d = 0.8$ is a strong effect.

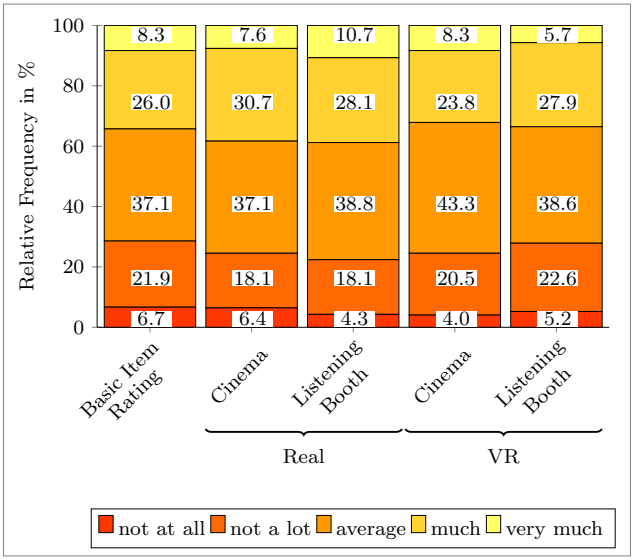

Fig. 11: Relative frequencies of the basic item ratings retrieved from the first session and the item ratings retrieved from the other four sessions.

Participants needed on average 6.7 min ($SD^2 = 2.2$) for the basic item session. A single real-world session took on average 6.9 min ($SD = 2.2$) and a single VR session on average 7.4 min ($SD = 1.1$). All item ratings retrieved from the participants are visualized by a frequency plot depicted in Figure 11.

Hypothesis I In order to test the first hypothesis ("The OLE ratings given in the VR environment do not differ from those given in the real-world experiment"), differences between the item ratings retrieved from the real-world sessions and from the VR sessions are investigated in more detail. The mean absolute difference is 0.51 stars ($SD = 0.62$) between item ratings given in the VR sessions and item ratings given in the real-world sessions. A Wilcoxon signed-rank test is applied to test whether the differences between the rating are statistically significant[3] (Wilcoxon, 1945). The Wilcoxon signed-rank test indicated that the ratings given in the VR environment were statistically significantly lower than ratings given in the real-world environment ($Z = -3.623, p =< .001$). However, the effect size of

[2] M = mean, SD = standard deviation, N = number of samples.

[3] The significance level α is set to 0.05 in this paper.

Coefficient	Estimate	p-value
basic item rating = 2	1.85	< .001
basic item rating = 3	3.56	< .001
basic item rating = 4	5.29	< .001
basic item rating = 5	6.82	< .001
env = VR	-0.07	.573
room = LB	0.20	.134
env = VR, room = LB	-0.31	.094
Threshold coefficients:		
	Estimate	
1 *Star* |2 *Stars*	-0.52	
2 *Stars*|3 *Stars*	1.97	
3 *Stars*|4 *Stars*	4.56	
4 *Stars*|5 *Stars*	7.10	
Maximum likelihood pseudo R^2: 0.45		
Cragg and Uhler's pseudo R^2: 0.48		

Table 2: Logit cumulative link model of the item ratings.

the differences is very low $r = \frac{Z}{\sqrt{N}} = -0.088$. A paired t-test results in the same outcome since the ratings were significantly different ($t(839) = -3.6, p =< .001$). Additionally, the weak effect is confirmed by Cohen's $d = -0.10$. To confirm the weak negative effect of the VR environment, a cumulative link model (CLM) was calculated. A CLM is a modification of a regression model for ordinal dependent variables (Agresti, 2002). The calculated CLM predicts item ratings based on the corresponding basic item ratings, the room and the environment. Additionally, the CLM checked for an interaction effect between room and environment. As one can see in Table 2, the VR environment has a very weak and non-significant effect on the item ratings. Furthermore, a negative interaction effect is detected for the VR environment in combination with the listening booth. This interaction effect is not significant for the chosen significance level α. In conclusion, according to our results, Hypothesis I must be rejected: OLE ratings obtained from the VR sessions are significantly lower than the ratings obtained from the real-world sessions.

Hypothesis II For testing the second hypothesis ("The OLE ratings that were given while being present in the (real-world) cinema do not differ from those given in the (real-world) listening booth"), the differences between the ratings given in different rooms are investigated. Overall, the mean of the absolute difference between the item ratings given in the cinema and given in the listening booth were 0.53 stars ($SD = 0.62$). A Wilcoxon signed-rank test ($Z = 1.937, p = .054$) indicated the difference between ratings given in the listening booth and in the cinema are not significant. The effect size of the differences is very low $r = 0.067$ (slightly positive effect for the listening booth). When analyzing the ratings of both rooms by a paired t-test, the difference turns out to be just statistically significant ($t(419) = 1.98, p = .048$). In accordance with the effect size r, Cohen's d also indicates a very weak effect of the listening room ($d = 0.078$). The cumulative link model (Table 2) confirms both results by showing a slightly positive effect of the listening booth. Moreover, in the cumulative link model, the effect size of the listening booth is not statistically significant different from zero. In conclusion, based on the results of the Wilcoxon signed-rank test and the cumulative link model, Hypothesis II is accepted. Based on the results of the paired t-test, Hypothesis II is rejected, but the effect size is very weak, thus practical importance of the differences is doubtful.

As mentioned in Section 1, we added some questions related to the room acoustics and the spatial perception of the participants. The relative frequency plots of the answers to these questions are shown in Figure 12. For each of these additional questions, a Wilcoxon signed-rank test was applied on the ratings given in the VR environment and on the ratings given in the real-world environment. Since the responses to these additional questions were not the main subject of interest, the independent variable room is not evaluated in detail. The answers to the question "How loud is this item?" indicated that the VR environment was statistically not significantly rated more quiet than the real-world environment ($Z = -1.988, p = .064, r = -.181$). According to the answers to question "How much reverb has this room?", participants indicated that they perceived more reverb in the VR environment than in the real-world environment ($Z = 0.482, p = .649, r = .044$). The stimuli of question "How far away do you perceive the sound?" were perceived further away in the VR environment than in the real-world environment ($Z = 1.028, p = .345, r = .094$). The answers to the question "How much do you like the bass of this item?" indicated that the bass was less liked in the VR environment than in the real-world environment ($Z = -2.986, p = .003, r = -.273$). The treble was also slightly less liked in the VR environment than in the real-world environment ($Z = -0.674, p = .600, r = -.061$).

After each VR session, the participants were asked whether they felt dizzy while taking part in the session. The participants answered 34 times with "no" (56.67 %), 19 times with "a bit" (31.67 %), and 7 times with "yes" (11.67 %).

4.3 Discussion

The experiment revealed some differences between the real-world environment and the VR environment. First

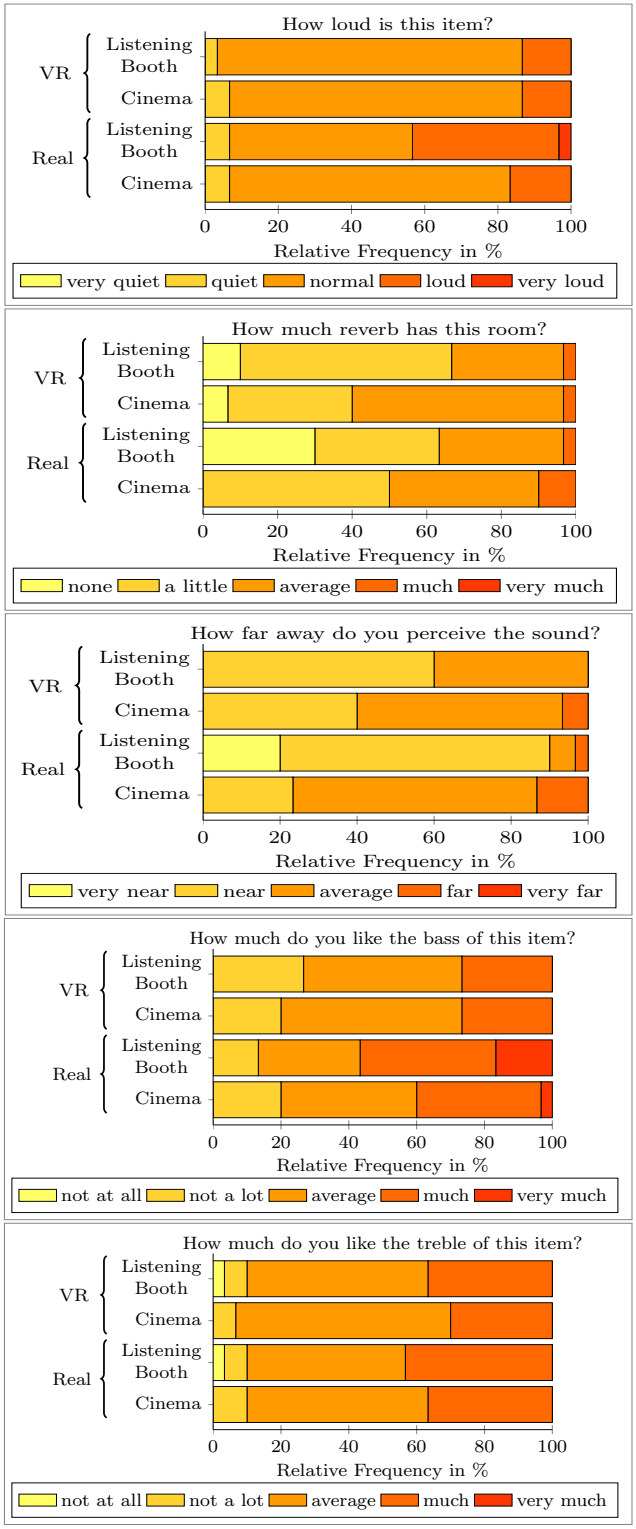

Fig. 12: Relative frequencies of the responses to the questions about room acoustics and spatial perception.

of all, participants needed more time in the VR sessions ($M = 7.4\,\text{min}$) than in the real-world sessions ($M = 6.9\,\text{min}$). Many participants reported that, especially at the beginning of the VR sessions, they spent a few moments to just look around and to examine the VR rooms. Spending some time just for examining the VR environment might not have a strong effect on the results of an experiment. The perfect outcome for the VR sessions would have been if the participants had not behaved significantly differently from the real-world sessions, meaning they would have needed approximately the same amount of time in both environments. This might have been achieved by additional training sessions in the same rooms that were later used in the VR sessions. Such additional training sessions would have allowed participants to familiarize themselves with the virtual rooms. These were not included because then additional training sessions should have been included for the real-world environment, too, in order to treat the two environments equally. Moreover, adding more training sessions would have been too time-consuming for the participants.

The first hypothesis had to be rejected as the item ratings retrieved from the VR environment turned out to be significantly lower than the ratings retrieved from the real-world environment. However, the effect size of the VR environment was found to be very weak. Especially, ratings given in the VR listening booth were lower than ratings given in the other sessions. Regarding this issue, participants reported that the instructions and control elements in the VR listening booth were harder to read than in the VR cinema, which might have had an influence on their overall enjoyment. One reason for the limited readability is that the VR listening booth was brighter and more colorful than the VR cinema, having the consequence that the "screen door effect" was more apparent. The VR cinema was a bit darkened, so the black lines separating the display's pixels were less perceivable. Furthermore, by being in a more colorful and brighter room, the limited resolution of the Oculus Rift becomes more perceivable. Another reason was that the virtual monitor of the listening booth scene was much closer to the participants than the stage of the VR cinema. Participants described the greater distance to the VR cinema stage as much more relaxing compared to the shorter distance to the monitor in the VR listening booth. Based on the weak effect sizes, we conclude that our results are in line with Bella's experiment (Bella, 2004), where the VR environment and the real-world environment also had a weak effect on the experiment's results. In summary, although the first hypothesis was rejected, we conclude that VR environments are suitable for experiments related to OLE since we found only minor differences between the VR and the real-world environment.

Based on results of a Wilcoxon signed-rank test and a cumulative link model, the second hypothesis had to be accepted since ratings given in the real-world listening booth were non-significantly higher than the ratings given in the real-world cinema. One reason for the slightly favored listening booth might be that the sound emitted from the surround channels was much better perceived in this room. In the cinema, the surround left and surround right channel are displayed by a large loudspeaker array where some loudspeakers were located at the front left and front right of the participant. Thus, the surround channels could be perceived as additional front channels. Moreover, the majority of the music excerpts were mixed as foreground-background mixes meaning that mainly ambient signal-parts were emitted by the surround loudspeakers. In addition, the loudness of the background was low compared to the foreground. Altogether, the music excerpts sounded more like stereo mixes in the cinema. As shown by a previous experiment, surround sound strongly enhances the OLE, which might be the reason for the lower ratings (Schoeffler et al., 2014b). However, music excerpts auditioned in the listening booth were not rated much higher than music excerpts listened to in the cinema. That surround sound is only slightly higher rated than stereo when rated in two different sessions has already been demonstrated by Schoeffler et al. (2014a).

The additional questions we asked to the participants were answered quite differently depending on the room and environment (see Figure 12). As already mentioned a few times before, analysis of these answers must be interpreted with great care since this part of the experiment is rather informal and these questions were added with the purpose of giving some indications for future experiments.

The loudness was equalized to the same level for all rooms and environments. The fact that participants would equally rate the apparent loudness was therefore more or less expected. However, in the real-world listening booth, participants rated the loudness as "loud" more often than in the other rooms. Unfortunately, no feedback from the participants addressed this issue, so we can only guess what the reasons could be. Loudness is perceived subjectively and dependent on many factors. For example, differences in perceived loudness between the real-world cinema and real-world listening booth can be explained by the findings of Mershon et al. (1981), who conducted an experiment to investigate the relationship between distance and perceived loudness. They found out that by increasing the distance but keeping the same loudness level, the loudness is perceived louder. In the listening booth, the speakers were located much closer to the participants than in the real-world cinema, so the loudness in the listening booth should be perceived to be quieter according to the findings of Mershon et al. However, our results are in contrast to those findings. Moreover, no other studies exist to our knowledge that investigated the loudness-to-distance relationship related to our scenario. Therefore, the differences in loudness ratings cannot be conclusively clarified. Another issue is that in the VR environment there is almost no difference between the answers, indicating that the loudness-to-distance relationship is not present in VR environments. In order to test this hypothesis, another VR experiment focusing on the effect of distance on the loudness must be performed.

The questions about the amount of reverb and the distance of the sound source were answered quite differently. Therefore, no conclusions can be drawn based on the answers to these two questions.

Participants differently answered the question about how much the bass was liked when listening to the *drums* stimulus. In particular, the bass was rated significantly higher in the real-world environment. The reason for this might be that the stimulus was reproduced by headphones in the VR environment and by loudspeakers in the real-world environment. Loudspeakers have the advantage of reproducing the bass of a stimulus more powerful than headphones. When participants rated the treble, there was no significant effect of the environment on the ratings. In contrast to reproducing the bass, headphones do not have such limitations with the treble.

5 Conclusion

A VR system was developed that allows creation of virtual scenes of experiments. The system renders the auditory stimuli by utilizing a set of BRIRs which is selected according to the user's head position. An experiment was implemented using the system, where ratings given in a VR environment were compared to ratings given in a real-world environment. In the experiment, participants rated the OLE of music excerpts while being in a cinema and a listening booth. For each room, the participants rated the music excerpts while being present in either the real-world or in the VR. Comparison of the results indicates that the ratings associated with the VR are slightly lower than the ratings retrieved from the real-world. In the real-world environment, music excerpts were rated slightly higher in the listening booth than in the cinema. In order to contribute to the validity of VR auditory experiments in general, future experiments must investigate other dependent perceptional variables such as sound quality, reverberation, loudness, and distance.

Acknowledgements The authors would like to thank Alexander Adami for taking pictures of the experiment apparatus and Marlene Röß for representing a participant in the pictures.

References

Agresti, A. *Categorial Data Analysis*. Wiley, 2nd edition, 2002. ISBN 0-471-36093-7.

Astheimer, P. What you see is what you hear-acoustics applied in virtual worlds. In *Virtual Reality, 1993. Proceedings., IEEE 1993 Symposium on Research Frontiers in*, pages 100–107, Oct 1993. ISBN 0-8186-4910-0.

Bella, F. Driving simulation in virtual reality for work zone design on highway: a validation study. In *The Second SIIV International Congress*, Florence, Italy, 2004.

Berkhout, A. J., de Vries, D., and Vogel, P. Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America*, 93(5):2764–2778, 1993.

Blauert, J. *Spacial Hearing, The Psychophysics of Human Sound Localization*. The MIT Press, Cambridge, Massachusetts, 1997. ISBN 978-262-02413-6.

Blauert, J. and Jekosch, U. A layer model of sound quality. *Journal of the Audio Engineering Society*, 60(1/2):4–12, 2012.

Bossard, C., Kermarrec, G., Buche, C., and Tisseau, J. Transfer of learning in virtual environments: a new challenge? *Virtual Reality*, 12(3):151–161, 2008.

Bowman, D. A. and McMahan, R. P. Virtual reality: How much immersion is enough? *Computer*, 40(7):36–43, July 2007.

Cakmakci, O. and Rolland, J. Head-worn displays: a review. *Display Technology, Journal of*, 2(3):199–216, Sept 2006.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates, 1988. ISBN 978-080-580283-2.

Cruz-Neira, C., Sandin, D. J., DeFanti, T. A., Kenyon, R. V., and Hart, J. C. The CAVE: Audio visual experience automatic virtual environment. *Commun. ACM*, 35(6):64–72, June 1992.

DeFanti, T. A., Dawe, G., Sandin, D. J., Schulze, J. P., Otto, P., Girado, J., Kuester, F., Smarr, L., and Rao, R. The StarCAVE, a third-generation CAVE and virtual reality optiportal. *Future Generation Computer Systems*, 25(2): 169–178, 2009.

European Broadcasting Union. Practical guidelines for production and implementation in accordance with EBU R 128 (version 2.0), 2011.

Frechaud, V. Gui3D, v. 1.11, 2013.

Fritz, C. O., Morris, P. E., and Richler, J. J. Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, 141(1):2–18, February 2012.

Garcia, G. Optimal filter partition for efficient convolution with short input/output delay. In *Proc. of the AES 113th Convention*, 2002.

Gardner, W. G. Efficient convolution without input/output delay. *J. Audio Eng. Soc*, pages 127–136, 1994. Preprint 3897.

Gilkey, R. and Anderson, T. *Binaural and Spatial Hearing in Real and Virtual Environments*. Taylor & Francis, 2014. ISBN 9781317780267.

Gorzel, M., Corrigan, D., Kearney, G., Squires, J., and Boland, F. Distance perception in virtual audio-visual environment. In *25th UK Conference of the Audio Engineering Society: Spatial Audio In Today's 3D World*, York, United Kingdom, 2012.

Gurusamy, K., Aggarwal, R., Palanivelu, L., and Davidson, B. R. Systematic review of randomized controlled trials on the effectiveness of virtual reality training for laparoscopic surgery. *British Journal of Surgery*, 95(9):1088–1097, 2008.

Hess, W. and Weishäupl, J. Replication of human head movements in 3 dimensions by a mechanical joint. In *Proc. of the International Conference on Spatial Audio (ICSA)*, 2014.

Kozak, J. J., Hancock, P. A., Arthur, E. J., and Chrysler, S. T. Transfer of Training from Virtual Reality. *Ergonomics*, 36(7):777–784, 1993.

Kuhlen, T., Assenmacher, I., and Lentz, T. A true spatial sound system for CAVE-like displays using four loudspeakers. In *Virtual Reality*, volume 4563, pages 270–279. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-73334-8.

Larsson, P., Västfjäll, D., and Kleiner, M. Perception of self-motion and presence in auditory virtual environments. In *in Proc. of Presence*, pages 252–258, 2004.

Lathi, B. P. and Green, R. A. *Essentials of Digital Signal Processing*. Cambridge University Press, 2014. ISBN 978-110-705932-0.

Le Callet, P., Möller, S., and Perkis, A. Qualinet white paper on definitions of quality of experience (version 1.1), 2012.

Lindau, A., Maempel, H., and Weinzierl, S. Minimum BRIR grid resolution for dynamic binaural synthesis. *The Journal of the Acoustical Society of America*, 123(5):3498–3498, 2008.

Loomis, J. M., Blascovich, J. J., and Beall, A. C. Immersive virtual environment technology as a basic research tool in psychology. *Behavior Research Methods, Instruments, & Computers*, 31(4):557–564, 1999.

Mershon, D. H., Desaulniers, D. H., Kiefer, S. A., Jr, A. T. L., and Mills, J. T. Perceived loudness and visually-determined auditory distance. *Perception*, 10(5):531–543, 1981.

Müller, S. and Massarani, P. Transfer-function measurement with sweeps. *J. Audio Eng. Soc*, 49(6):443–471, 2001.

Novo, P. Auditory virtual environments. In *Communication Acoustics*, pages 277–297. Springer Berlin Heidelberg, 2005.

Palomäki, H., Kalle J.and Tiitinen, Mäkinen, V., J.C. May, P., and Alku, P. Spatial processing in human auditory cortex: The effects of 3d, ITD, and ILD stimulation techniques. *Cognitive Brain Research*, 24(3):364–379, 2005.

Pearson, J. L. and Dollinger, S. J. Music preference correlates of jungian types. *Personality and Individual Differences*, 36(5):1005–1008, 2004.

Prince, W. F. A paradigm for research on music listening. *Journal of Research in Music Education*, 20(4):445–455, 1972.

Psotka, J. Immersive training systems: Virtual reality and education and training. *Instructional Science*, 23(5-6):405–431, 1995.

Pysiewicz, A. On the validity of web-based auditory perception experiments [master thesis]. Master's thesis, TU Berlin, 2014.

Rose, F. D., Attree, E. A., Brooks, B. M., Parslow, D. M., Penn, P. R., and Ambihaipahan, N. Training in virtual environments: transfer to real world tasks and equivalence to real task training. *Ergonomics*, 43(4):494–511, 2000.

Rumsey, F., Zielinski, S., Kassier, R., and Bech, S. Relationships between experienced listener ratings of multichannel audio quality and naive listener preferences. *The Journal of the Acoustical Society of America*, 117(6):3832–3840, 2005.

Sanchez-Vives, M. V. and Slater, M. From presence to consciousness through virtual reality. *Nature Reviews. Neuro-*

science, 6(6):332–339, May 2005.

Sandel, T. T., Teas, D. C., Feddersen, W. E., and Jeffress, L. A. Localization of sound from single and paired sources. *The Journal of the Acoustical Society of America*, 27:842–852, 1955.

Schoeffler, M. and Herre, J. About the impact of audio quality on overall listening experience. In *Proceedings of Sound and Music Computing Conference*, pages 48–53, Stockholm, Sweden, 2013.

Schoeffler, M. and Herre, J. Towards a listener model for predicting the overall listening experience. In *Proc. of Audiomostly 2014*, Aalborg, Denmark, 2014a.

Schoeffler, M. and Herre, J. About the different types of listeners for rating the overall listening experience. In *Proceedings of Sound and Music Computing Conference 2014*, Athens, Greece, 2014b.

Schoeffler, M. and Hess, W. A comparison of highly configurable CPU-and GPU-based convolution engines. In *Audio Engineering Society Convention 133*, San Francisco, USA, 2012.

Schoeffler, M., Edler, B., and Herre, J. How much does audio quality influence ratings of overall listening experience? In *Proc. of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, pages 678–693, Marseille, France, 2013a.

Schoeffler, M., Stöter, F.-R., Bayerlein, H., Edler, B., and Herre, J. An experiment about estimating the number of instruments in polyphonic music: A comparison between internet and laboratory results. In *Proceedings of 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, 2013b.

Schoeffler, M., Adami, A., and Herre, J. The influence of up- and down-mixes on the overall listening experience. In *Proc. of the AES 137th Convention*, Los Angeles, USA, 2014a. Preprint 9140.

Schoeffler, M., Conrad, S., and Herre, J. The influence of the single/multi-channel-system on the overall listening experience. In *Proc. of the AES 55th Conference on Spatial Audio*, Helsinki, Finland, 2014b.

Schoeffler, M., Stöter, F., Edler, B., and Herre, J. Towards the next generation of web-based experiments: A case study assessing basic audio quality following the itu-r recommendation bs.1534 (MUSHRA). In *1st Web Audio Conference*, Paris, France, 2015.

Schröder, D., Wefers, F., Pelzer, S., Rausch, D., Vorländer, M., and Kuhlen, T. Virtual reality system at RWTH Aachen University. In *Proceedings of the International Symposium on Room Acoustics*, Sydney, New South Wales, Australia, 2010. Australian Acoustical Society, NSW Division.

Schuemie, M. J., van der Straaten, P., Krijn, M., and van der Mast, C. A. Research on presence in virtual reality: A survey. *CyberPsychology & Behavior, and Soc. Networking*, 4(2):183–201, 2001.

Seeber, B. U. and Fastl, H. On auditory-visual interaction in real and virtual environments. In *Proceedings of the 18th International Congresses on Acoustics*, pages 2293–2296, Kyoto, Japan, 2004.

Silzle, A., Strauss, H., and Novo, P. IKA-SIM: A system to generate auditory virtual environments. In *Audio Engineering Society Convention 116*, May 2004. Preprint 6016.

Stanney, K. Realizing the full potential of virtual reality: Human factors issues that could stand in the way. In *Proceedings Virtual Reality Annual International Symposium '95*, pages 28–34. IEEE Comput. Soc. Press, 1995.

Stanney, K. M., Mourant, R. R., Kennedy, R. S., and Literature, A. R. O. T. Human factors issues in virtual environments: A review of the literature. *PRESENCE*, 7:327–351, 1998.

Steuer, J. Defining virtual reality: Dimensions determining telepresence. *Journal of Communication*, 42(4):73–93, December 1992.

Stockham, T. G., Jr. High-speed convolution and correlation. In *Proceedings of the April 26-28, 1966, Spring Joint Computer Conference*, pages 229–233, New York, NY, USA, 1966. ACM.

Sveistrup, H., McComas, J., Thornton, M., Marshall, S., Finestone, H., McCormick, A., Babulic, K., and Mayhew, A. Experimental studies of virtual reality-delivered compared to conventional exercise programs for rehabilitation. *Cyberpsy., Behavior, and Soc. Networking*, 6(3):245–249, 2003.

The OGRE Team. OGRE game engine, v. 1.9.0, 2013.

Torger, A. and Farina, A. Real-time partitioned convolution for ambiophonics surround sound. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics 2001*, pages 195–198, 2001.

Väljamäe, A., Larsson, P., Västfjäll, D., and Kleiner, M. Auditory presence, individualized head-related transfer functions, and illusory ego-motion in virtual environments. In *Proc. of 7th Annual Workshop on Presence*, 2004.

van Dam, A., Laidlaw, D. H., and Simpson, R. M. Experiments in immersive virtual reality for scientific visualization. *Computers & Graphics*, 26(4):535–555, August 2002.

Västfjäll, D. The subjective sense of presence, emotion recognition, and experienced emotions in auditory virtual environments. *CyberPsychology & Behavior*, 6(2):181–188, 2003.

Vora, J., Nair, S., Gramopadhye, A. K., Duchowski, A. T., Melloy, B. J., and Kanki, B. Using virtual reality technology for aircraft visual inspection training: Presence and comparison studies. *Applied Ergonomics*, 33(6):559–570, 2002.

Welch, N. and Krantz, J. H. The world-wide web as a medium for psychoacoustical demonstrations and experiments: Experience and results. *Behavior Research Methods, Instruments, & Computers*, 28(2):192–196, 1996.

Werner, S. and Siegel, A. Effects of binaural auralization via headphones on the perception of acoustic scenes. In *Proc. of 3rd International Symposium on Auditory and Audiological Research (ISAAR)*, Nyborg, Denmark, 2011.

Werner, S., Liebetrau, J., and Sporer, T. Audio-visual discrepancy and the influence on vertical sound source localization. In *Proc. of 4th Int. Workshop on Quality of Multimedia Experience, QoMEX*, pages 133–139, Melbourne, Australia, 2012.

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, December 1945.

Witmer, B. G., Bailey, J. H., Knerr, B. W., and Parsons, K. C. Virtual spaces and real world places: Transfer of route knowledge. *Int. J. Hum.-Comput. Stud.*, 45(4):413–428, 1996.