Binaural walk-through scenarios with actual self-walking using an HTC Vive

Annika Neidhardt, Niklas Knoop
Institut für Medientechnik, TU Ilmenau, 98693 Ilmenau, Deutschland
Email: annika.neidhardt@tu-ilmenau.de, niklas.knoop@tu-ilmenau.de

Introduction

In the field of virtual reality, interactive exploration has become a common requirement. Nowadays, inexpensive tracking devices allow precise tracking of orientation and position even at home. Hence, the scene can be controlled and explored by actual self-motion. Head mounted displays allow a dynamic reproduction of visual information while sound is usually provided via headphones. To create a convincing experience, a plausible dynamic reproduction of spatial sound is required.

In the study presented in this paper, an HTC Vive system is used to realize position-dynamic binaural synthesis. The listeners could walk through a virtual acoustic environment themselves. In particular, the cases of walking towards a sound source and walking past a sound source are investigated. A listening experiment was conducted to evaluate the perception of these walk-throughs considering different virtual acoustic scenes. Simulated scenes as well as scenes created from measurements were taken into account. To focus on auditory perception, no visual information of the source positions or the virtual room were provided. Only some basic visual cues, that are necessary for the orientation within the setup, were available.
With the results, some factors for limited plausibility are revealed. Furthermore, the need of appropriate methods to evaluate the perception of self-translation within virtual acoustic environments is discussed.

Interactive binaural audio scenes

Creating virtual acoustic environments (VAEs) that allow listeners to walk through the scene themselves are not only interesting for gaming applications. Such VAEs can be useful e.g. for industrial auralisations or to investigate the human auditory perception in dynamic scenarios. This requires a detailed knowledge of the potentials, as well as the limits and drawbacks of state-of-the-art systems.
A popular method to reproduce virtual acoustic scenes in real-time via headphones is the crossfading between orientation- and position-dependend binaural room impulse responses (BRIRs). A list of technical issues, that may degrade the quality of the auditory illusion, is known. So far only few studies on the perception of a walk-through in dynamic binaural scenes have been conducted.
In this paper, we present a study investigating the plausibility of different virtual scenes created with quick crossfading between binaural room impulse responses. This is a quite basic approach. Reproduction methods based on interpolation or even extrapolation, e.g. from measured data would be of interest. Knowing more details about the capability of this simple approach to provide a plausible auditory illusion, will be fundamental for further studies on more efficient algorithms.

Real-time rendering of binaural audio

The dynamic reproduction of binaural audio in real-time is usually based on partitioned convolution. The efficiency of an implementation depends on various technical parameters. In the dissertation of Wefers [1] different approaches are presented and discussed in detail.
The system used in this study is based on the uniformly partitioned convolution realised with the overlap-save approach. This method was implemented in Python with pyAudio as the fundament for the audio reproduction. The filters are switched with quick binaural crossfading (within few samples). As a result the user actually listens to the same filters for a certain section of the way.

Plausibility of virtual acoustic scenes

One first goal of a perceptual evaluation is to find out, whether a convincing auditory illusion of the desired virtual environment is created.
Kuhn-Rahloff [2] defined the term *plausibility* as an agreement with the inner reference, which is the result of the individual listening experience. But the inner reference does not only vary between different people. It also seems to underlie certain dynamics. Therefore designing listening experiments that rely on a comparison to the inner reference is challenging.
Another difficulty is that many people do not listen carefully to their environment. Thus, they might not have an appropriate inner reference, or worse: they could have a wrong one.
Lindau and Weinzierl [3] suggested a method to test for plausibility by asking the participants if they hear a simulated or a real scene. According to this suggestion plausibility is given, when the participants cannot distinguish the simulation from the real scene. Consequently this method requires a corresponding real scene for each virtual scene to be tested. Especially for fictive scenes this is not possible. But fictive scenes can still be plausible.

This study attempts to investigate the plausiblity of virtual acoustic environments without providing any real scene for comparison. As discussed above, this approach comes along with some challenges. But the results will be an important step towards the establishment of an appropriate test method.

Creating scenes allowing self-translation

For this experiment, six different scenes were generated. Each of the scenes consisted of a set of BRIRs corresponding to 9 different positions along a straight line in a distance of 25cm to each other (fig. 1). The angular resolution was 4° for the azimuth. Changes in elevation were not taken into account.
Two different sound source positions were chosen. In one case, the source was placed in the front, hence the listener would walk towards it. In the other case, it was positioned at the side, enabling the listener to walk past it. In both cases, the closest distance between source and listener was 1.25m. Fig. 2 shows the setup.
These two scenerios were studied in three different room settings. First, BRIRs were captured along a line of nine points in a real room for both source positions. Additionally, shoebox-versions of a similar and a clearly more reverberant room were simulated.

Measurement of BRIRs

A G.R.A.S. Kemar 45BA with large ears was used to measure the BRIRs of the room, where experiment was carried out. The listening lab (complying ITU-R BS11.16-2) has a size of 8.4m× 7.6m×2.8m, $V = 179m^3$ and a reverberation time $T_{60} = 0.28s$ (broad band). Two loudspeakers Genelec 1030A were used as sound sources according to the room setup shown in fig. 2.

Simulation of BRIRs

With MCRoomSim [4] a shoebox-shaped equivalent of the listening lab was created. The directivity data of a Tannoy-V6 loudspeaker provided by the toolbox and spherical HRTF data set of a Neumann KU100 dummy head [5] were used to model source and receiver in this scenario.
A second room with the same size and equivalent source and listening positions was simulated. The absorption coefficients were reduced to increase reverberation time to $T_{60} = 1.2s$.

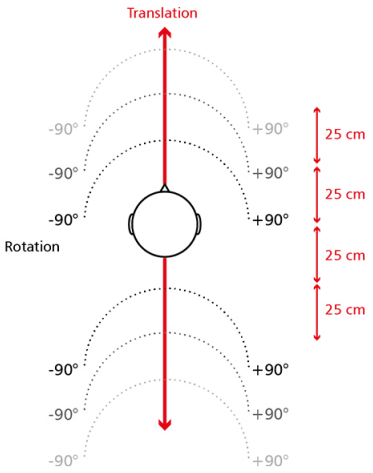

Figure 1: Arrangement of BRIRs for the dynamic binaural reproduction in the first experiment, second: full 360° [6]

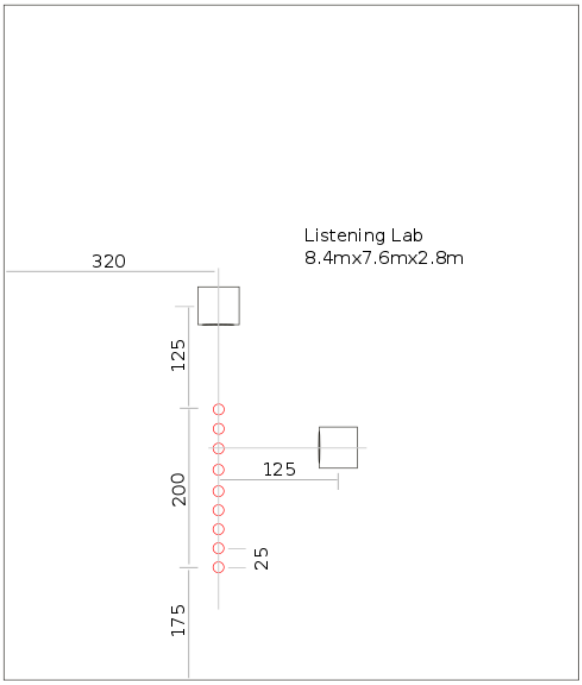

Figure 2: Basic room setup for all scenes used in the study

Technical setup

The setup consisted of an audio signal processing unit, the tracking module of the HTC Vive with its two infrared cameras and the sensors within the display device itself. With this module, the position and orientation of the user's head were tracked and the related data was sent via OSC (Open Sound Control) messages to the real-time rendering unit, which selected the corresponding binaural FIR filter and applied it to the signal.

With the *Room Scale* option of the HTC Vive, a certain area in a room can be set up for moving around. When getting close to the border, a visual feedback is given to the user in terms of an appearing blue grid. The area is visually marked by a rectangle on the ground. The translation line was setup along one of the lines indicating the origin. So the participants could actually see the line on the ground, which the should not leave while listening. Additionally they saw a neutral grid, which did not provide any information about the virtual room or the source positions.
For the audio reproduction *STAX SR-Lambda Professional new* headphones were used in combination with headphone compensation filters.

Table 1: Overview of tested scenes, all following the same geometrical setup shown in fig. 2

Scene	HATS	Room
MeasLab Front	KEMAR 45BA	T_{60}=0.3s
MeasLab Side	KEMAR 45BA	T_{60}=0.3s
SimLab Front	Neumann KU100	T_{60}=0.3s
SimLab Side	Neumann KU100	T_{60}=0.3s
SimRev Front	Neumann KU100	T_{60}=1.2s
SimRev Side	Neumann KU100	T_{60}=1.2s

First Experiment

Nine people with an average age 28.1 years, three of them female, took part in this experiment. Neither of the participants had experienced a virtual self-walk-through with headphones before. Each participant was invited for two sessions. At the beginning of both sessions, the subjects could listen to three different scenes for getting used to exploring the virtual acoustic scene on their own. Afterwards, in a test design adapted from the Absolute Category Rating (ACR) [7], the plausibility of the six scenes (Tab. 1) had to be rated on a scale from 0-100, with 100 as the value for maximal plausibility. Each participant evaluated each scene once per session in a randomized order. A short piece of music played by a solo sax was used as a test stimuli.

Results and observations

Fig. 3 provides an overview of the plausibility ratings. The values are spread over the whole range of the scale. In most cases no normal distribution could be found. Furthermore, big differences of the ratings in the first and the second session were observed. Within these differences no certain trend could be found. The participants seemed to be very unconfident regarding their ratings.

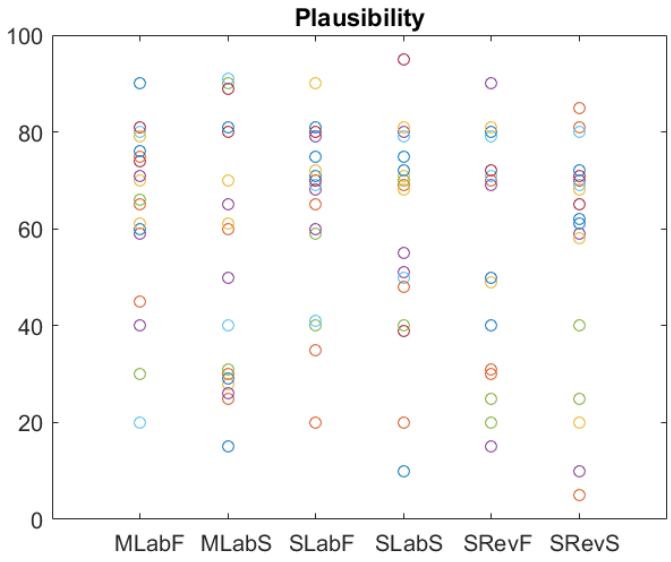

Figure 3: Results of first experiment: Rating *Plausibility* on a scale from 0 to 100 without a reference or direct comparison to a real or simulated scene

For each of the scenes averages were slightly above 50. The ratings for the scenes with the virtual loudspeaker in the frontal position show a slightly more positive trend. Furthermore, in the individual ranking no clear patterns could be found.
If a participant did not choose the highest rating, he was asked to describe the reasons for the degradation.
The following list provides an overview:

- localising sound source internally or close to head
- unstable/moving sound source
- reduced localizability/ diffuse localization (especially in the more reverberant virtual room)
- loudness progress a bit confusing (less/higher change than expected)

Second Experiment

Yes-No paradigm

Lindau and Weinzierl [3] assessed the plausibility of a virtual acoustic environment using a Yes-No paradigm. For their experiment specific measurements of BRIRs with a dummy head wearing headphones were conducted. Applying that method to translational movements requires a huge measurement effort. Additionally, fictive scenes like the second simulated room cannot be investigated in the suggested way, because an equivalent real scene is not available. Therefore, the second experiment is designed using the Yes-No paradigm in a different way.
12 people with an average age 30.3 years, three of them female, were asked to answer the following question with *Yes* or *No*:

- Did you get the impression of walking towards/past a sound source?
- Would you call this experience a plausible illusion of a sound source?

Further changes to the first experiment

During the first experiment participants reported, that they were confused in the beginning, but after some movements the scene started to make sence to them. A listener might need some time to understand a scene, maybe even with real scenes. Therefore, in the second experiment participants had to walk up and down the line at least once before rating the scene.
Furthermore, according to reports from the listeners, it is confusing if the reproduction is limited to a certain angle region, e.g. from -90° to +90° like in the first experiment. Participants carried out some exploration movements and suddenly the illusion broke down because of a confusing progress of the sound field while turning around. Hence, it seems to be important to provide a full 360° reproduction, if the quality of the illusion is investigated. In the second experiment this aspect was taken into account.
Each of the scenes was provided once with dry male speech (8min excerpt of an audio book) and once with mono music (one channel of a stereo pop song) for each participant. The order was shuffled for each session.

Results and observations

The two columns in fig. 4 give an overview of the answers to both questions for the different test items. Although all scenes were evaluated in a randomized order, the scenes which differ only by the source signal show similar results. A clear preference of the scenes based on measured BRIRs can be observed. A longer T_{60} seems to reduce the drawbacks of the simulated BRIRs.

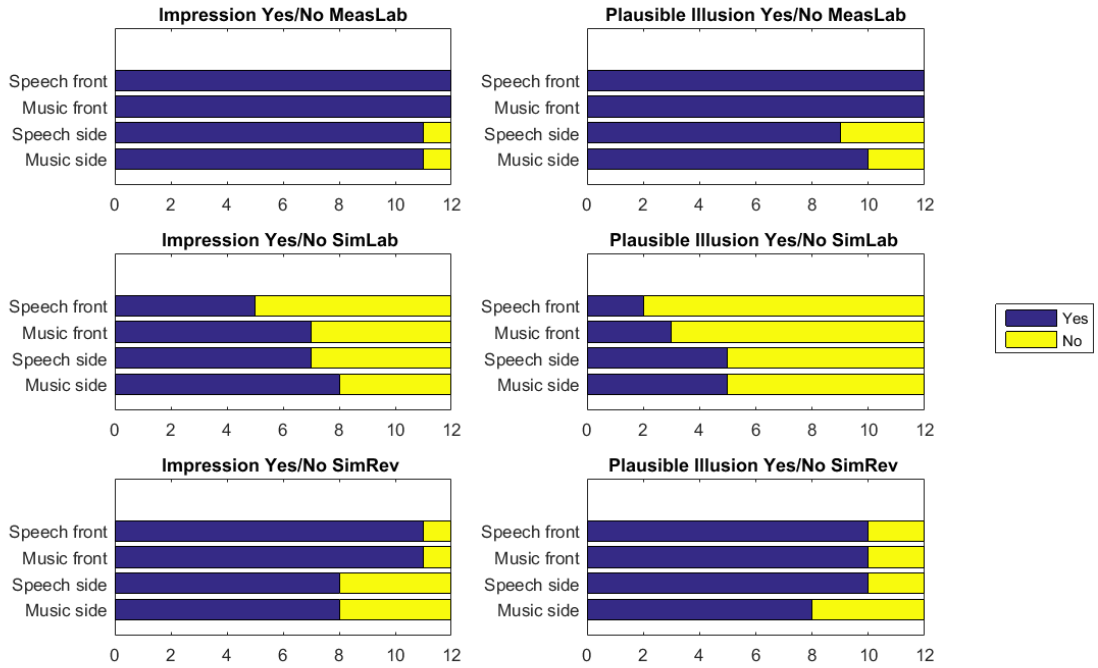

Figure 4: Results of second experiment based on Yes/No questions: Answers to "Did you get the impression of walking towards/past a sound source?" (left), Answers to "Would you call this experience a plausible illusion of a sound source?" (right)

Summary and conclusion

This paper describes a basic realization of VAEs allowing the listener to walk through the scene and presents two experiments to study the perception of such scenes.
The test method used in the second experiment, brought up perceptual differences between the different scenes. All participants stated, that a plausible illusion of walking towards a virtual loudspeaker was achieved, when measurements along a 25cm grid were used. On the other hand, a simulation with very similar acoustic properties caused some perceptual issues, that reduced the quality of the illusion. Limitations in externalisation and sound source stability were reported.
Scenes built on the same BRIRs show similar results. It is concluded, that the test method is suitable to evaluate plausibility. Further verification is necessary, because the relatively long exploration time per scene does not allow many repetitions, which are needed to achieve a good statistical reliability.
In the future it would be of interest to rate or measure plausibility with a finer resolution. The limited knowledge about variances and individual differences of the inner reference remain a challenge.
Furthermore it was observed, that in the simulation the audible change of distance due to walking did not suit the expected change. The phenomenon was described with expressions like "not enough change in loudness/coloration" or "sound source is following while walking away". The exact reason causing this phenomenon needs to be identified in further experiments.
The results of the second experiment are part of a bigger study. Besides plausibility further attributes like coloration, externalisation, stability of the sound source and sound field continuity were evaluated by the participants. The remaining data will be analysed and published soon.

Acknowledgements

Thanks to Alexander Raake and his group for providing the HTC Vive for this study. Thanks to all the participants for their effort and their interest. This work has been supported by the Free State of Thuringia (2015FGR0090) and the European Social Fund.

References

[1] Frank Wefers. *Partitioned convolution algorithms for real-time auralization.* PhD thesis, RWTH Aachen, 2014.

[2] C. Kuhn-Rahloff. *Prozesse der Plausibilitätsbeurteilung am Beispiel ausgewählter elektroakustischer Wiedergabesituationen.* PhD thesis, TU Berlin, 2011.

[3] A. Lindau and S. Weinzierl. Assessing the Plausibility of Virtual Acoustic Environments. *Acta Acustica united with Acustica 98(5):804-810*, 2012.

[4] A. Wabnitz, N. Epain, C. Jin, and A. van Schaik. Room acoustics simulation for multichannel microphone arrays. In *Int. Symp. on Room Acoustics, ISRA, Melbourne, Australia*, 2010.

[5] B. Bernschütz. A Spherical Far Field HRIR/HRTF Compilation. In *39. Jahrestagung für Akustik, Meran*, 2013.

[6] N. Knoop. Orientierung in virtuellem Raum mit bewegl. Avatar. Master's thesis, TU Ilmenau, 2016.

[7] Int. Telecommunication Union (ITU). Rec. ITU-T P.910 Subjective video quality assessment methods. Technical report, Genf, Schweiz, 2008.