Comparison of Different Methods to Create an Interactive Augmented Auditory Reality Scenario Using Sparse Binaural Room Impulse Response Measurements

Stephan Werner[1], Annika Neidhardt[1], Florian Klein[1], Karlheinz Brandenburg[1,2]

[1] *Technische Universität Ilmenau, Electronic Media Technology Group, 98693 Ilmenau,*
E-Mail:{stephan.werner, annika.neidhardt, florian.klein, karlheinz.brandenburg}@tu-ilmenau.de
[2] *Fraunhofer Institute for Digital Media Technology, 98693 Ilmenau.*

Introduction

An auditory illusion of a spatial acoustic environment can be created with the help of existing spatial audio systems. Psychoacoustically adequate ear signals can be created by using binaural synthesis approaches. The creation of a plausible auditory illusion is shown coherently in a wide range of research projects. But the occurrence of such a plausible auditory illusion depends on an adequate technical realization and on several context dependent quality parameters like congruence between synthesized scene and the listening environment or individualization of the technical system. A scenario with room acoustic divergence yields a clear decrease of perceived externalization. Room acoustic divergence can occur if synthesis methods of binaural room impulse response (BRIRs) are used.

This contribution presents a binaural synthesis system, which adds a virtual sound source to the real room. The realization is based on BRIRs measured with a sparse distribution in that room. To provide a certain area of action for the listener, a dense grid of BRIRs is generated using energy- and time-based BRIRs synthesis approaches. The evaluation of the system includes the quality features overall impression, localization stability, and externalization under test condition with translation and head movement of the listener. For comparison purposes, a fully measured dataset is included in the test.

Auditory Augmented Reality

The term augmented reality addresses a subset of the so-called reality-virtuality continuum described by Milgram et al. [1]. Figure 1 classifies this subset in the mixed reality approaches. A real-world environment is augmented with perceptional events which are caused by the synthesis of virtual objects. The presented auditory augmented reality is a technical mediated reality which adds virtual audio objects at defined positions in a real room. The virtual objects are simulated by using a binaural synthesis system. The assumed aim of this system is to create a plausible auditory illusion for the listener.

The technical system has to fulfil the user´s expectation on the presented and augmented environment. The challenge is to minimize perceptional conflicts resulting from divergences between the perceived and desired nature of the constructed reality in our cognitive system. These divergences can yield a negative effect on quality describing perceptional features like localization, externalization, timbre, spaciousness, or others.

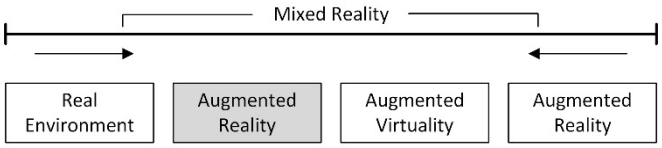

Figure 1: The reality-virtuality continuum after Milgram et al. [1]; gray area addresses the research field in this paper.

The augmented reality scenario enriches an office-like room with virtual audio objects at fixed positions in the room. A listener can move within this room. The ear signals are synthesized depending on his position and head orientation.

Figure 2 gives a sketch of the auralized room. The virtual audio objects are marked as numbers 1 and 2. The range of motion of the listener is limited in this system by the marked grid positions.

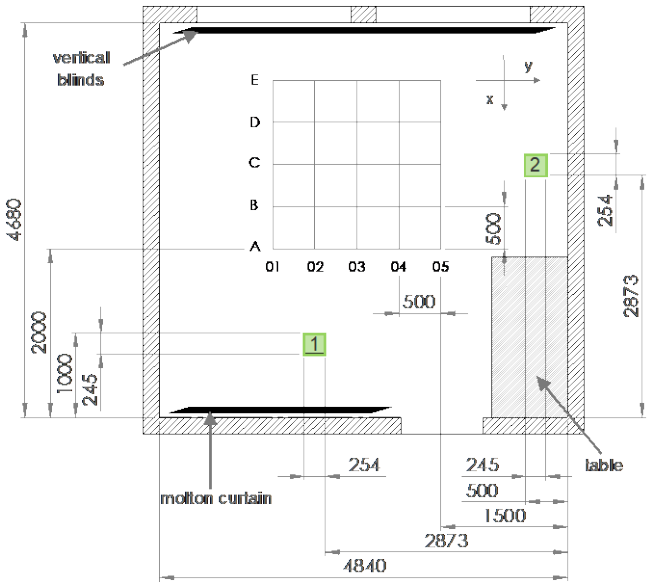

Figure 2: Sketch of the auralized room and the possible range of motion of the listener as grid in the middle of the room. The virtual audio object positions are marked as numbers 1 and 2; figure from Mittag [2].

Binaural Synthesis System

A binaural synthesis system is used to synthesize the ear signals. The system includes a head and position tracking of the listener. The system uses BRIRs for the left and the right

ear from each auralized point of the area in Figure 2. Only the head yaw and the horizontal position of the listener is used for auralization. The geometric resolution of the grid is 0.5 m x 0.5 m. The resolution of the yaw angle is set to 5°. It is assumed that these limitations yield to perceptional effects in localization of the virtual audio objects.

The BRIRs are either recorded with a head and torso simulator (dummy head KEMAR) at each of the 25 grid positions, or synthesized by the methods which are described in the next section.

BRIR Synthesis

Three methods for creation of the needed BRIRs are investigated in this contribution. A full measurement of the 25 grid positions is used as a reference in the evaluation. This system is called **25M**.

Figure 3 gives a sketch of the different measured positions and synthesis approaches.

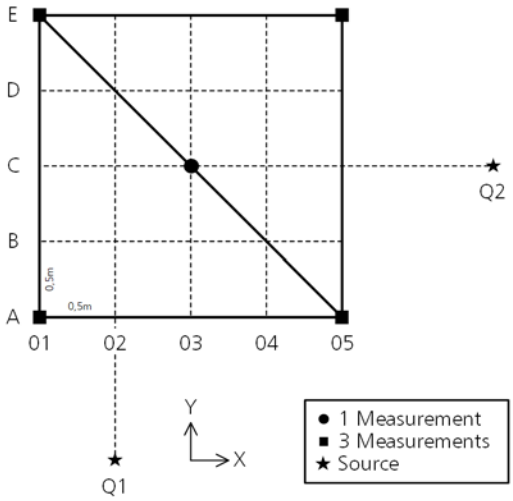

Figure 3: Grid for auralization with measured positions at each grid point for the reference system (25M) and the positions used for the approaches 3M and 1M.

The approach **3M** uses BRIRs from three measured positions to synthesize new BRIRs at the grid points within the spanned triangle. The measured BRIRs are split into the direct sound (until 1.5 ms after the direct sound), the early reflections (until the perceptional mixing time; here approx. 80 ms), and the late reverberations. The direct sound and the late reverb of the closest measured BRIRs are taken and adapted in the intensity depending on the new synthesized grid position. The early reflections of all three measured BRIRs are interpolated. Frequency components below 1.5 kHz are interpolated linearly in time-domain and a weighting is applied. The magnitude spectrum of the frequency components above 1.5 kHz are interpolated in frequency domain. The phase information is taken from the measured BRIR closest to the synthesis position. The yaw orientations of the new BRIRs are realized by selecting the corresponding direct sound of the measured BRIRs.

The approach **1M** uses one measured set of BRIRs. The BRIRs are also split in a direct part, reverb part, and early reflections. The yaw orientation is realized by selecting the corresponding direct sound of the measured BRIRs. The intensity of the direct sound and reverb sound is adapted according the new synthesized grid position. The early reflections are stretched or compressed in time-domain depending on the change of the Initial Time Delay Gap for the new position.

A much more detailed description of the used methods is given in the work from Mittag [2, 3] and Füg [4, 5].

The quasi real-time realization of the filter convolution and integration of a position and head pose tracking is done by the pyBinSim tool described in [6]. The active tracking device from a HTC Vive is used to catch the position and orientation of the listener´s head.

Test Design

A listening test is performed to investigate the influence of the different BRIR synthesis methods and of the auralized grid positions on quality features. The quality features are externalization of the auditory event, the ability to perceive a direction of the auditory event (localization ability), the stability of the perceived position of the auditory event in the room (localization stability), and the overall impression. The used rating scales are shown in Figure 4.

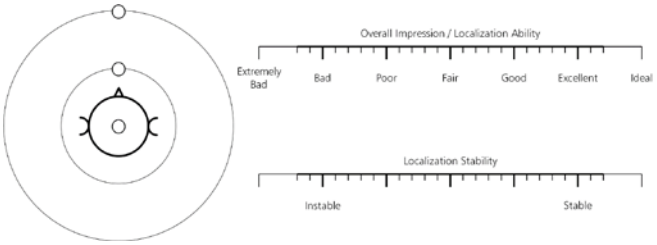

Figure 4: Rating scales for the different quality features in the listening test; left circle is for externalization with inside the head, close to the head, and external perceived event.

Eighteen test persons with a mean age of 29 years (SD=9 years) participate in the listening test. Eleven participants report experience in perceptional testing as test persons. Only seven report experience in binaural synthesis listening.

A speech signal of a male English speaker is used as audio signal. The signal, with a duration of approx. 30 s, is looped during the exploration of the acoustic environment through the test person.

Several paths within the grid are used for evaluation of critical test conditions. Figure 5 shows the used paths for the test and training. A training is conducted to familiarize the test persons with the used binaural synthesis system and the rating procedure. The test persons have to rate the quality features for three training paths and for auralization of the sound source 2 (see Figure 5 right). Source 1 is used for the main quality test. The test person has to start at a designated position and has to move with a self-chosen walking speed to the end of the path. If the end is reached, he can turn the direction or walk backwards until a qualified rating concerning the features is possible.

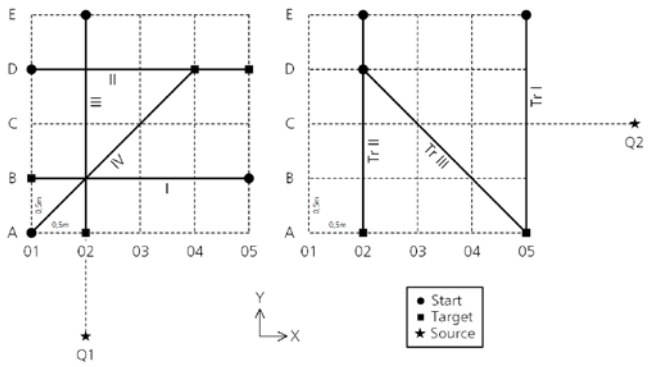

Figure 5: Moving paths of the listeners in the test; left=test paths, right=training paths.

Four paths and source 1 are used for the quality rating. Path I is assumed as critical because of the close distance to the audio object and a rapid change of the relative angle between head and source position. Path II includes farer distances and less rapid change of angles. Path III covers far and close distances. Furthermore, it is intended that this path includes most direct front and back head orientations to the sound source. Path IV includes the most critical jumps of the filter functions for the different grid positions.

Ratings

The ratings are counted as index for externalization. The index shows the frequency of externalized auditory events (outer circle in Figure 4 left) in ratio to the overall number of ratings. The ratings for the other features are presented as boxplots on the used scales. The shown scale points from 0 to 6 correspond to the main ticks from left to right of the used scales in Figure 4.

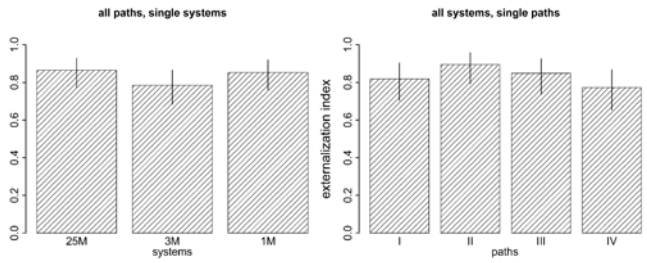

Figure 6: Externalization as index with 95% binominal conf. interval; left=systems, right=paths.

The externalization shows no significant differences ($p<.05$) between the used BRIR synthesis systems or evaluation paths (Figure 6). The system 3M and the paths I and IV are rated in tendency with lower externalization compared to the other systems and paths. This is probably caused by the closer distances to the audio object and/or by the synthesis approach used in 3M. The early reflections are a weighted interpolation between three recording points. This cause a confusion of the early reflection pattern compared to measured BRIRs at these positions. The confused pattern may yield to less externalization.

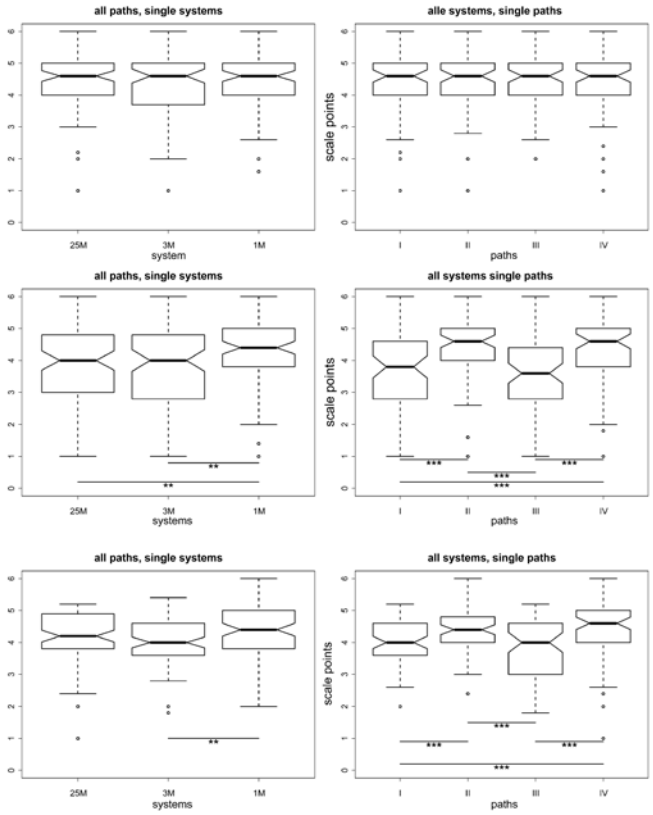

Figure 7: Top: localization ability; Middle: localization stability; Bottom: overall impression; ratings on the scale shown in Figure 4 with main scale points from left (0) to right (6); **significant difference at $p<.05$, ***significant difference at $p<.01$.

The rating for the ability to perceive a direction of the auditory event is shown in the top of Figure 7. No significant differences ($p<.05$) between the synthesis systems or evaluation paths are observed. A slightly bigger interquartile distance for system 3M compared to the other systems is visible. This is probably also caused by the confusion of the early reflection patterns. The overall good localization ability is the results of the usage of the unchanged (except of an intensity adaptation) direct sound of the recorded BRIRs.

The stability of the perceived position of the auditory event is shown in the middle of figure 7. Significant differences are visible between the paths and the systems. The path I and III are downgraded compared to the other paths. These paths are assumed as most critical for perceived jumps between the BRIRs of the different grid points (path I) and the low yaw angle resolution especially for frontal directions (path III).

Furthermore, the 1M system reached significant higher ratings compared to the other systems. This is a surprise. It is hypothesized that the usage of only one BRIR set and the ITDG shaping does not create big confusions in the early reflection patterns of the synthesized BRIRs. The new patterns are smooth changes of the recorded ones. Although the synthesized reflection pattern differs in some acoustic parameters from a measured BRIR the new BRIRs seems to be very viable for the investigated quality feature.

The bottom part of Figure 7 shows the ratings for the overall impression of the auditory augmented reality. The systems

and the paths are rated as “good” (scale points > 4). Significant differences are again visible for the paths I and III because of probably the same reason as for the feature localization stability. The system 3M is downgraded in its overall impression compared to the other systems, probably because of the confusion of the reflection patterns.

Conclusion

This contribution gives an insight in an auditory reality system using binaural synthesis. The BRIRs are synthesized by using three or only one measured BRIRs. The investigated systems yield proper quality ratings for the features externalization, localization ability, localization stability, and overall impression.

The system using one BRIR measurement position reaches similar quality ratings than the reference system using measurements from all 25 positions in the room. The system using three BRIR measurement positions is downgraded compared to the reference system and the system using one measurement position. Furthermore, the used spatial resolution of the auralization grid becomes critical for closer distances because of jumps in the synthesized direction of the audio object. The resolution of the yaw angle becomes most critical for frontal head orientations to the audio object. An adaptive change of the spatial resolution is one next step in the proposed system. This includes an increase of the spatial density in the synthesized BRIR positions and an interpolation of the direct sound to increase the yaw resolution.

We thank all test participants in our tests and the students in the course “Advanced Psychoacoustics” at TU Ilmenau. We thank Christina Mittag and Simone Füg for their master projects on related topics. This work and the underlying project is funded by the Free State of Thuringia and the European Social Fund.

References

[1] Paul Milgram et al., “Augmented Reality - A class of displays on the reality-virtuality continuum”. In: SPIE-Telemanipulation and Telepresence Technologies 2351 (1994).

[2] Christina Mittag, Stephan Werner and Florian Klein, “Development and Evaluation of Methods for the Synthesis of Binaural Room Impulse Responses based on Spatially Sparse Measurements in Real Rooms”. In: 43. Jahrestagung für Akustik, DAGA, Germany, 2017.

[3] Christina Mittag, “Entwicklung und Evaluierung eines Verfahrens zur Synthese von binauralen Raum-impulsantworten basierend auf räumlich dünnbesetzten Messungen in realen Räumen“. Master Thesis, Technische Universität Ilmenau, Electronic Media Technology Group, Germany, 2016.

[4] Simone Füg, Stephan Werner, and Karlheinz Brandenburg, “Controlled Auditory Distance Perception using Binaural Headphone Reproduction – Algorithms and Evaluation”, In proceeding of: VDT Int. Convention, 27. Tonmeistertagung, Germany, 2012.

[5] Simone Füg, “Untersuchungen zur Distanzwahr-nehmung von Hörereignissen bei Kopfhörer-wiedergabe”. Master Thesis, Technische Universität Ilmenau, Electronic Media Technology Group, Germany, 2012.

[6] Annika Neidhardt, Florian Klein, and Thomas Köllmer, "Flexible python tool for dynamic binaural synthesis applications", 142nd Convention of the Audio Engineering Society (AES), Germany, 2017.