Development and Evaluation of Methods for the Synthesis of Binaural Room Impulse Responses based on Spatially Sparse Measurements in Real Rooms

Christina Mittag[1,2], Stephan Werner[1] and Florian Klein[1]

[1] *Institute for Media Technology, Ilmenau University of Technology, Ilmenau, Germany*
[2] *now with Fraunhofer IIS, Erlangen, Germany*
Email: christina.mittag@iis.fraunhofer.de, stephan.werner@tu-ilmenau.de, florian.klein@tu-ilmenau.de

Introduction

The motive for this research is the development of an interactive audio system for a moving listener. Such a system could for example be used in a museum to give exhibits a voice and such realize a new, individual and immersive experience for visitors. It shall be capable of auralizing a real room enriched with auditory perceivable objects by using measured binaural room impulse responses (BRIRs). In order to reduce the expenditure of time and costs for the realization of this system, the number of required BRIRs measured in the real room needs to be minimized, without losing sound quality or plausibility of the sound scene. Three algorithms are developed which synthesize BRIRs in a spatial context. The synthesis methods use measurements from one to three positions in the room to generate new BRIRs at defined positions. They make use of adjustment of the perceived distance and spatial interpolation of the BRIRs. By using the synthesized BRIRs for the auralization, spatial subsampling could be avoided while the number of measurements is reduced. The synthesized BRIRs are compared to measured ones with respect to their technical applicability by investigating the direct-to-reverberant energy ratio (DRR). Furthermore, a listening test is conducted to analyze the sound quality and the externalization of the synthesis results in comparison to the measured BRIRs. The tests display a satisfying sound quality, no significant differences in perceived externalization and dependencies between the quality of the synthesis results and the combinations of source positions, synthesis positions and measurement positions.

Background

Interactive audio systems for moving listeners already exist, but they usually use simulated BRIRs [1]. But simulated audio scenes often lack plausibility and measured BRIRs could provide more convincing auditory results. Furthermore, clear evidences are available which show that congruence between the synthesized room and scene yields to a more plausible spatial perception if a binaural headphone system is used [2]. The drawback is the amount of measurements that have to be done for such a system. Several approaches have been made in the last years to reduce the amount of measurements for different kinds of datasets used for binaural synthesis. Savioja et al [3] calculated head related transfer functions (HRTFs) for points on a reference sphere by bilinear interpolation from the four nearest measured HRTFs on the sphere. Algazi [4] developed a technique called Motion Tracked Binaural (MTB) sound for capturing, recording and reproducing spatial sound. He uses a circular array of microphones on a sphere with the diameter of an average head to capture the sound. When playing back the sound via headphones the movement of the listener's head in the horizontal plane is tracked and the headphone signals are interpolated from the microphone recordings closest to the ear position. Füg [5] and Sass [6] used different techniques to change the distance impression of BRIRs and Kearney [7] simulated changes in source movement by interpolation of room impulse responses (RIR). The approach presented in this paper synthesizes BRIRs in a spatial context of a real room by choosing suitable BRIRs out of a dataset and combining Füg's distance adaption technique [5] with spatial interpolation.

Method

360°-BRIR measurements are performed at 25 positions in an empty office room at TU Ilmenau (V=73 m³, RT60=1.15 s) using a KEMAR head and torso simulator (45BA with ears KB0065/6). The measurement positions are defined by a rectangular measurement grid of 2 x 2 m², also called listening area, and took place in 5°-steps for two sound sources. The room with measurement grid can be seen in Figure 1. The resulting dataset consists of 3600 BRIRs (one BRIR consisting of left and right channel); 1800 BRIRs for each source [8,9]. Two Geithain MO-2 active studio monitors are used for the BRIR measurements.

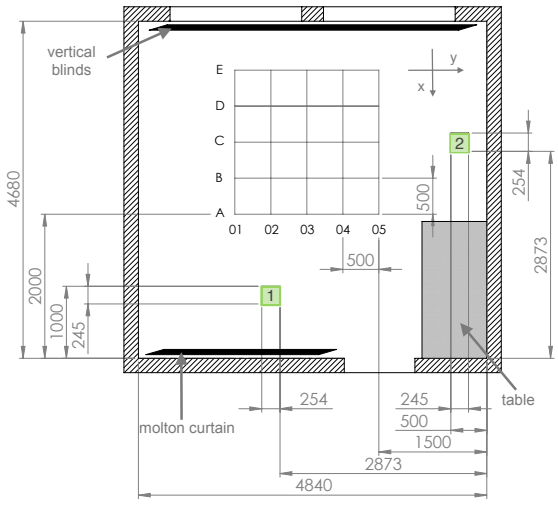

Figure 1 – Test room with measurement grid; small squares with numbers indicate the loudspeaker positions

Algorithm oM – one Measurement

The first developed synthesis algorithm oM calculates a new BRIR at a given position and for a given angle of incidence of sound in the room by using only one user-determined 360°-BRIR measurement [8]. It choses the BRIR with the correct angle of incidence from the dataset and adjusts its

distance to the new distance from the synthesis position in the room to the source by using an ITDG-Shaping algorithm by Füg [5]. The ITDG-Shaping algorithm adapts the initial time delay gap and the energy of the BRIR to the new synthesis distance. Errors could appear if the reflections in the BRIR, that do not match the new position in the room, disturb the plausibility of the perceived sound scene.

The second and third algorithm makes use of adjustment of distance and spatial interpolation of three measurements, which form a triangle, to synthesize a BRIR [8]. Depending on the chosen position in the measurement triangle (see Figure 2), weights for the interpolation are calculated. This happens similarly to the weight calculations in Vector Base Amplitude Panning (VBAP) [10]. With the given measurement positions, synthesis position, source position and angle of incidence of sound the distances between those positions and the angles for the BRIR selection are calculated. Only the BRIRs with the correct angle of incidence at the measurement points are considered for the interpolation. The first step for both algorithms is the ITDG-Shaping and the energy adjustment to the new synthesis distance for the chosen BRIRs. By doing this it can be assured that the first reflection is approximately at the same sample in all three BRIRs, which is important for the interpolation. Then the direct sound of the BRIRs is separated from the spatial part. The direct sound is not interpolated in the synthesis process to preserve its direction indicating property, but is taken from the measurement position, which is closest to the synthesis position. The interpolation of the spatial parts differs for the two algorithms.

Algorithm wS – weighted Sum

Algorithm wS adds up the weighted spatial BRIR parts to obtain the synthesized spatial part. This step is followed by an energy adjustment of both direct sound and spatial part to adapt the DRR to the synthesis distance. Finally the two parts are merged and the pre-delay is adapted.

Algorithm MTB – Motion Tracked Binaural

Algorithm MTB performs a frequency dependent interpolation similar to a method used for the MTB microphone [11]. The BRIRs are separated into direct sound, early reflections and the reverberant part. The splitting point between early reflections and reverberant part results from the perceptual mixing time of the room of 51 ms [12]. In algorithm MTB only the early reflections are interpolated, whereas the reverberant part is taken from the closest measurement position. The frequency content below 1500 Hz is interpolated in the time domain, the high frequency content above 1500 Hz is interpolated in the frequency domain. The energy of the three BRIR parts is adjusted to the synthesis distance before the parts are merged and the pre-delay is adapted.

Subjective Quality Measures

The synthesized BRIRs are evaluated in two listening tests for several test scenarios. The scenarios include two positions in the room (B02 and C04), three angles of incidence of sound at those positions (0°, 30°, 120°) and measurement positions that have two different distances to the synthesis positions (D01, C03 – for Algorithm oM) and form synthesis triangles of two different sizes (A01, A05, E05, C03, D01 – for Algorithm wS and MTB). The test scenarios are visualized in Figure 2. The BRIRs for those scenarios are synthesized with each developed algorithm and result in six systems under test (each algorithm combined with small and big test scenario). As a reference the measured BRIRs at the same positions as the positions under test are chosen. The BRIRs are convolved with two different dry audio signals: a male voice and a saxophone. The test signals are equalized using non-individual headphone transfer functions (HPTFs) from the KEMAR. A Stax Lambda Pro New headphone is used for playback. The listening tests take place in the same room, in which the BRIRs were measured. The test persons are seated in the middle of the room. The listening position does not coincide with one of the tested synthesis positions.

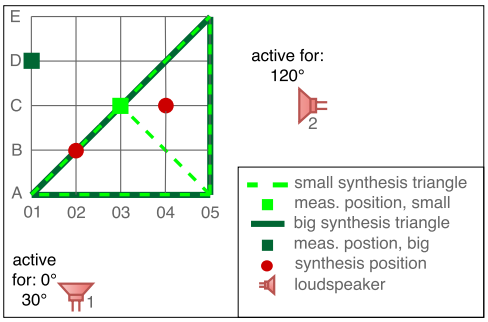

Figure 2 – Test scenarios for two synthesis positions C04, B02; green dark and bright lines and squares represent the measurement positions and synthesis triangles (formed by measurement positions at corners of triangles). Loudspeaker 1 is active for directions 0°, 30°; loudspeaker 2 is active for 120°.

To evaluate the sound quality of the synthesized BRIRs, a test following the ITU-R BS.1116 standard is performed [13]. The standard is selected to investigate small impairments between the reference and the synthesized signals. The test persons have to judge the difference between synthesis, hidden reference and reference on a scale from 1 (very big difference / very disturbing) to 5 (no difference audible). A number of 36 test items per synthesis position has to be assessed (6 systems, 2 test signals, 3 angles of incidence). Position C04 is evaluated by four female and 16 male test persons (average age: 28 years). Position B02 is evaluated by seven male test persons (average age: 29 years). Both groups had to take part in a training with half of the test items.

The test persons rated the perceived externalization of the auditory event in the second listening test part by using a graphical user interface (GUI) [2]. The GUI can be seen in Figure 3. It shows a top-down view of the listener's head and three dashed circles, which represent three different externalization zones and are defined as follows: a) "The auditory event is located inside the head of the listener or is very diffuse" (internal), b) "The auditory event is external but really close to the listener's head" (close), c) "The auditory event is external and easy to localize" (external). The test persons have to assign the perceived auditory event to one of the small circles, according to the externalization definitions. Anchors, that should represent an in-head

localization by using only the direct sound part of the BRIRs, are used in this test [14]. A training with 30 of the 96 test items (7 systems, 2 test signals, 3 angles of incidence, 2 positions, 12 anchors) took place before the listening test to familiarize the test persons with the GUI and the perceived externalization. There are 17 male and five female listeners in the test group (average age: 28 years). 14 people indicated that they are familiar with binaural synthesis.

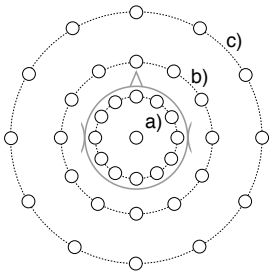

Figure 3 – Graphical User Interface for second listening test part with three externalization zones.

Results

The analysis of the sound quality ratings is based on quartiles, 95%-confidence intervals and statistical tests. The ratings for the hidden reference are subtracted from the ratings for the systems under test for evaluation. No significant differences are found for the two test signals (male voice and saxophone).

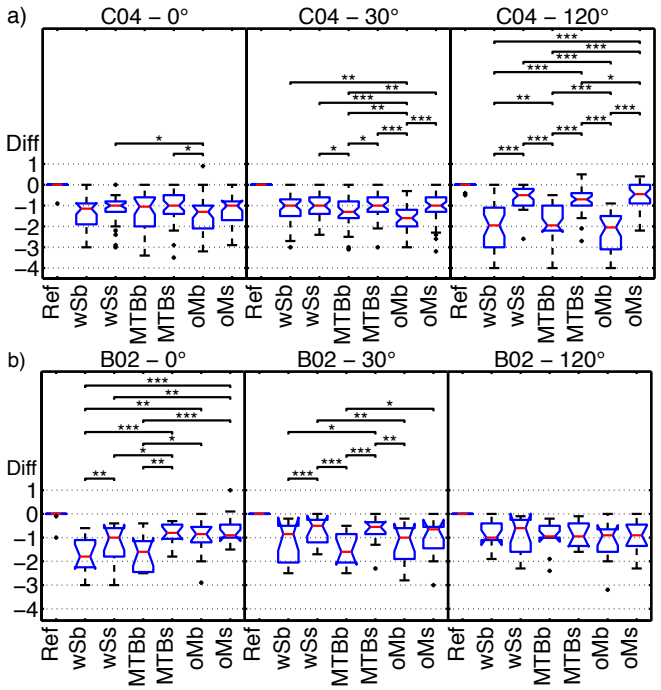

Figure 4 – Results of quality test at position a) C04 and b) B02 for directions 0°, 30°, 120° as boxplots; significant differences without reference: ∗=p≤0.1, ∗∗=p≤0.05, ∗∗∗=p≤0.01; Ref=Reference, wS=weighted Sum, MTB=Motion Tracked Binaural, oM=one Measurement, s=small, b=big.

Figure 4 shows the differences between the systems' and hidden reference's ratings for the two synthesis positions C04 and B02 for angles of incidence of sound of 0°, 30° and 120°. The test persons are in most of the cases able to identify the reference signal. All systems show non-disturbing but significant differences compared to the reference, which are not separately displayed in the figures. At position C04 for 0° and 30° and position B02 for 120° all systems got rated 0.5 to 1.5 points worse than the reference and show only a few significant differences. Striking significant differences between the systems can be seen for position C04 for 120° and position B02 for 0° and 30°. The small-systems wSs, MTBs and oMs are rated significantly better than the big-systems. This leads to the conclusion that the use of small synthesis triangles that are placed (symmetrically) in front of the active sound source as well as a smaller distance between synthesis and measurement position results in a better synthesis quality. Differences between the systems and the reference are caused by coloration and localization differences. These are for example caused by the synthesis of the reverberant part by interpolation.

Figure 5 shows the results of the second listening test part, divided according to the two tested distances between synthesis position and source position (1.5 m and 2.25 / 2.55 m). The externalization index EI is calculated by dividing the number of external ratings by the total number of test items.

$$EI = \frac{N_{ext}}{N_{ext} + N_{close} + N_{int}} \qquad (1)$$

The results show a significant higher EI for the bigger distance between synthesis and source position. The externalization index is lowest for an angle of incidence of 0° and highest for 120° for both distances between synthesis position and active sound source. The anchor is rated least external for all conditions. All in all the systems under test do not show a different behavior compared to the reference system, except for system oMb, which is rated significantly worse than the reference several times. The externalization indexes for different distances are comparable with results from other listening tests using measured BRIRs from several distances [2].

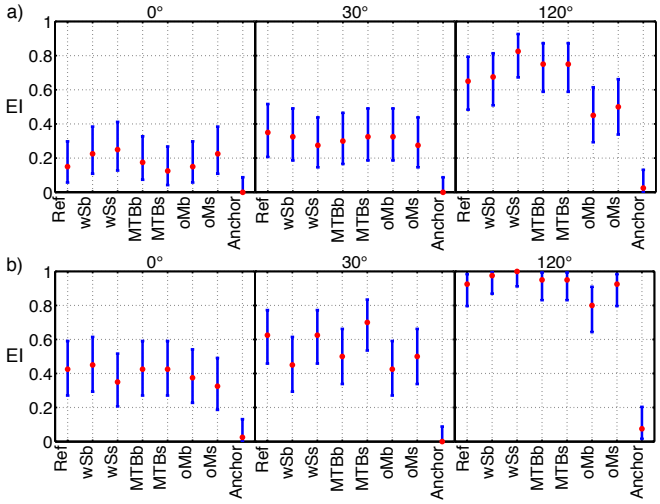

Figure 5 – Externalization indices with 95% confidence intervals for distances a) 1.5m and b) 2.25 m / 2.55 m for directions 0°, 30°, 120°; Ref=Reference, wS=weighted Sum, MTB=Motion Tracked Binaural, oM=one Measurement, s=small, b=big.

Objective Quality Measures

As an objective quality measure the DRR is chosen for investigation. It is a highly distance determining value [15].

Conformity of the DRR of the measured BRIRs with the DRR of the synthesized BRIRs would increase the potentiality that both BRIRs produce the same distance perception [16]. The DRRs are compared over the whole listening area for one angle of incidence by building the difference value DRR_{Diff} between the reference measurement and the synthesis at every position and interpolate in between.

$$DRR_{Diff} = DRR_{Ref} - DRR_{Syn} \qquad (2)$$

An example for the evaluation of the left channels of the BRIRs can be seen in Figure 6 for algorithms wSb/MTBb, wSs/MTBs, oMb and oMs for an angle of incidence of 0°.

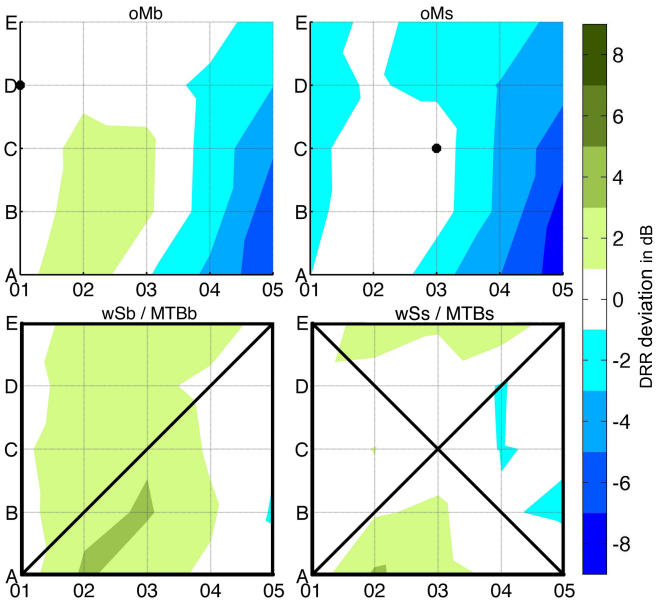

Figure 6 – DRR evaluation for direction 0° for the algorithms with active source 1, 1 m in front of A02; top= oMb and oMs , black points show the used measurement positions; bottom= wSb/MTBb, wSs/MTBs, big black lines show the used synthesis triangles and measurement positions (corners of triangles).

For the algorithms wS and MTB the DRR deviations are around 2 dB, so very small and inside the limit of the just noticeable difference that is given by Larsen [17]. For algorithm oMb and oMs deviations appear on the edge of the listening area. It can be concluded that most of the deviations arise because of the directional radiation of the used speakers, which is not considered in the algorithms.

Conclusions

The results show that all of the developed algorithms can be used for the spatial synthesis of BRIRs. The sound quality of the synthesis results was rated throughout satisfying and there was no increase in inside-head localization found compared to the measured BRIRs. Only the algorithm oMb showed significantly worse results in the externalization test with up to 0.2 scale points less than the measured reference. Dependencies on the used constellations of source position, synthesis position and measurement positions can be seen. It can be stated that the best synthesis quality is obtained when using small synthesis triangles / small distances between measurement position and synthesis position. The algorithms wS and MTB, which use synthesis triangles, perform best when the triangles are positioned symmetrically in front of the active sound source. It has to be investigated if the algorithms wS and MTB yield benefits compared to the algorithm oM. Looking at the objective quality measure the algorithms manage to approximate the DRR of the measured BRIRs but are not capable of approximating the directional radiation of the sound sources.

Acknowledgment

The authors thank the test persons for participation in the listening test. This work is supported by grants of the Deutsche Forschungsgemeinschaft (DFG Grant BR 1333/14-1) and by Thüringer Aufbaubank (2015FGR0090) and the European Social Fund.

References

[1] Zimmermann, A. and Lorenz, A.: “LISTEN: a user-adaptive audio-augmented museum guide“, User Model User-Adapt Inter, 19, pp. 389-416, DOI 10.1007/s11257-008-9049-x, 2008.

[2] Werner, S., Klein, F., Mayenfels, T., and Brandenburg, K.: “A Summary on Acoustic Room Divergence and its Effect on Externalization of Auditory Events”, of 8th Int. Conference on Quality of Multimedia Experience (QoMEX), Portugal, 2016

[3] Savioja, L. et al: “Creating interactive virtual acoustic environments”, J. Audio Eng. Soc. (47)9, 1999

[4] Algazi, V.R. et al: „Motion-Tracked Binaural Sound“, J. Audio Eng. Soc. (52)11, 2004

[5] Füg, S.: „Untersuchungen zur Distanzwahrnehmung von Hörereignissen bei Kopfhörerwiedergabe“, Master Thesis, Technische Universität Ilmenau, Germany, 2012

[6] Sass, R.: „Synthese binauraler Raumimpulsantworten“, Master Thesis, Technische Universität Ilmenau, Germany, 2012.

[7] Kearney, G. et al: “Towards Efficient Binaural Room Impulse Response Synthesis”, EAA Symp. on Auralization, Finland, 2009

[8] Mittag, C.: “Entwicklung und Evaluierung eines Verfahrens zur Synthese von binauralen Raumimpulsantworten basierend auf räumlich dünn besetzten Messungen in realen Räumen”, Master Thesis, Technische Universität Ilmenau, Germany, 2016.

[9] Mittag, C., Böhme, M., Werner, S., and Klein, S.:” Dataset of Binaural Room Impulse Responses at Multiple Recording Positions, Source Positions, and Orientations in a Real Room”, to be published in Proc. of the 43rd annual convention for acoustics, DAGA, Germany, 2017.

[10] Pulkki, V.: “Virtual sound source positioning using vector base amplitude panning”, J. Audio Eng. Soc., 45(6), 1997

[11] Lindau, A. and Roos, S.: “Perceptual evaluation of discretization and interpolation for motion-tracked binaural (MTB) recordings”, VDT International Convention, 26. Tonmeistertagung, Germany, 2010.

[12] Lindau, A. and Kosanke, L.: “Perceptual evaluation of physical predictors of the mixing time in binaural room impulse responses”, 128th AES Convention, 2010

[13] Recommendation ITU-R BS.1116-3 (2/2015) “Methods for the subjective assessment of small impairments in audio systems including Multichannel Sound Systems”. International Telecommunication Union, Radio communication Assembly.

[14] Begault, D. R., Wenzel, E. M.: “Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source”, J. Audio Eng.Soc.,49,pp.904-916, 2001.

[15] Blauert, J.: “Spatial Hearing - The Psychophysics of Human Sound Localization”, MIT Press, 1997.

[16] Werner, S., Klein, F., and Sporer T., “Adjustment of the Direct-to-Reverberant-Energy-Ratio to Reach Externalization within a Binaural Synthesis System”, AES Conf. on Audio for Virtual and Augmented Reality, Los Angeles, CA, USA, 2016.

[17] Larsen, E. et al: “On the minimum audible difference in direct-to-reverberant energy ratio”, J. Acoustical Soc. America, Vol. 36, pp. 259-291, 1992