Evaluation of a perceptually optimized room-in-room reproduction method for playback room compensation

Julian Grosse
Acoustics Group, Cluster of Excellence "Hearing4All", University of Oldenburg, Germany.
Email: julian.grosse@uni-oldenburg.de

Steven van de Par
Acoustics Group, Cluster of Excellence "Hearing4All", University of Oldenburg, Germany.

Summary
In sound reproduction it is usually desired to reproduce a sound source as accurate as possible to achieve a natural and realistic sound representation of the sound source. A perceptually motivated sound reproduction approach was suggested by [Grosse, van de Par, IEEE, 2015] where only the perceptually most relevant cues of a sound field in the recording room are considered. These perceptually relevant cues are the coloration of the sound source, the interaural cross correlation (IACC), which is linked to the listener envelopment, and the reverberation time. Based on this the direct and reverberant sound are rendered over spatially distributed loudspeakers in the playback room while optimizing the direct and reverberant sound field such that the considered cues of the recording room coincide with the reproduced cues in the playback room. A MUSHRA-test showed rather good results for the proposed optimization (listening room compensation, sweet-spot robustness) compared to a simple room-in-room or multichannel reproduction in terms of the overall sound quality. This follow-up study considers several important perceptual aspects and general problems which will occur when a sound source is rendered in a playback room. A subjective listening test is presented where the focus is on perceptual attributes like the perceived timbre, realism and plausibility of various alternative approaches to sound reproduction compared to the perceptually optimized room-in-room reproduction.

PACS no. 43.60.Dh, 43.60.Uv, 43.55.Hy, 43.60.Cg

1. Introduction

Various different approaches and goals to sound reproduction could be considered. We will consider the challenge to reproduce a sound source, which was recorded in a specific recording room, in a reverberant playback room in such a way, that the listener has an authentic and plausible representation of the sound source signal including the room acoustical properties of the recording room. Breebaart et al. [1] suggested an approach in the context of low-bit audio coding with which a nearly perfect spatial reconstruction is possible by conserving only the three most important binaural cues, interaural time delay, interaural level difference and interaural cross correlation (IACC). Based on this observation, a perceptually motivated sound reproduction approach was suggested by [2] where only a limited set of perceptually relevant parameters were considered for optimizing the sound reproduction. These parameters were the coloration of the sound source, the reverberation time T_{60}, and the interaural cross correlation [3]. These cues were measured and optimized at the listening positions in the recording and playback room. General problems which will occur when reproducing a recorded sound source in a reverberant playback room are that the listener will hear a convolutive mix of both room impulse responses of the playback and recording room. Due to this convolution, not only an extended reverberation time can be observed [4]. The spectral coloration, represented in the spectral standard deviation, will change by a factor of $\sqrt{2}$ as compared to the 5.6 dB that is theoretically predicted [5]. The coloration problem could be solved by equalizing the transfer function between the loudspeakers and listener in the playback room. However, it will not be possible to create a flat transfer function for both the

direct sound and the reverberant sound field. In addition, it can be expected that the IACC will differ between the recording and playback room. This can lead to a perceived narrowing of the phantom source between the loudspeaker due to the reduction of lateral reflections which are existing in the recording room. A solution for the reproduction of the lateral reflections is to excite the reverberant sound field of the playback room with additional rear loudspeakers, e.g. in a multi channel setup. To excite the diffuse field with additional rear-speakers can solve the IACC problem. In addition, the rear speakers can be used to optimize the reverberant sound field provided that the sound played over the rear loudspeakers is delayed relative to the sound on the front loudspeakers.
A previous study [2] showed in a headphone experiment, that the authenticity of the perceptually motivated reproduction is rather high as compared to several other rendering methods.
In this paper, a study is presented where the perceptual optimized sound field is presented over loudspeakers. The advantage of loudspeaker reproduction as compared to headphone reproduction is that listeners will listen with their own HRTFs within an actual reproduction room. In this study it will be compared with several conventional reproduction methods. In the listening test several (spatial) attributes like reverberance, naturalness, apparent source width (ASW), listener envelopment, perceived distance to the sound source, and coloration are investigated.

2. Method

The following Section describes the technical setup of the recording and playback room for the stereo playback, the multi channel playback, as well as the perceptual optimization which were used in this study.
In principle it should be possible to have an accurate reproduction of the physical sound field at the listeners ears. This can be achieved by optimize the reproduced sound field in such a way that:

$$h(t)_{\mathrm{ref}} = h(t)_{\mathrm{pr}}$$

where $h(t)_{ref}$ is the binaural room impulse response from source to listener in the recording room and $h(t)_{\mathrm{pr}}$ is the binaural room impulse responses (BRIR) of the full reproduction chain consisting of recording and playback room. This approach will, however, not be very robust against small changes e.g. in the listener positions. Therefore we aim for reproducing the sound field in a perceptually accurate manner. For this we optimize the auditory transfer function:

$$\mathrm{ATF}_{\mathrm{ref}} = \mathrm{ATF}_{\mathrm{pr}}$$

The ATF represents the energy in each auditory filter (with a critical ERB-spaced bandwidth [6]) at the listeners ear-drum and is modelled with a Gammatone-Filterbank [7]. Figure 1 shows the technical setup in

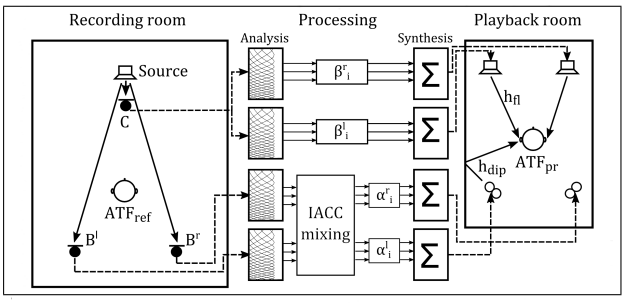

Figure 1. Measurement and reproduction setup. The recording room is shown on the left side including an artificial head that measures the perceptually relevant parameters. In addition a artificial head is placed in the playback room (shown on the right side) to reproduce the sound recorded in the recording room. Microphone C records the direct sound close to the sound source. This signal is optimized with a Gammatone-Filterbank in such a way that the direct sound in the playback room and the recording room have the same spectral shape as measured on the listeners positions. The microphones $B^{(l,r)}$ are located at two distant positions in order to record the diffuse sound field. These signals are rendered via two dipole loudspeakers in the playback room in such a way that the diffuse field is similar in the recording room and playback room as measured at the listeners position. An inter-channel cross-mixing allows for optimizing the IACC in the playback room.

the recording and playback room in detail. An artificial head is placed at the location of the listener in front of the sound source. The binaural impulse response, measured on the artificial head ref is used to determine the desired parameters. In addition, an artificial head is placed in a 60° stereo triangle in the playback room to be able to optimize the perceptual parameters at this listening place. The parameters are the coloration, which is represented as the spectral energy distribution, the interaural cross correlation and the reverberation time (T_{60}). Based on the binaural room impulse responses, the sound field at the listeners location can be decomposed in a direct and reverberant path:

$$h(t)_{\mathrm{ref}} = h(t)_{\mathrm{ref},d} + h(t)_{\mathrm{ref},rev}$$

Based on the knowledge that the direct and reverberant sound field is reproduced separately over the front and rear loudspeakers, $h_{\mathrm{ref},d}$ and $h_{\mathrm{ref},rev}$ can be used to control the amount of direct and reverberant sound, respectively. On basis of the decomposed direct and reverberant paths an auditory transfer function (ATF) can be derived to obtain the spectral energy distributions of both paths. The interaural cross correlation is derived from the whole BRIR $h(t)_{\mathrm{ref}}$ and is represented as the normalized-cross correlation. The third parameter, reverberation time is derived from the mean of the left and the right ears BRIR. The signal C records the direct sound with a single microphone close to the sound source in order to have

a rather dry signal of the sound source. In practice the signal C contains a small amount of early and lateral reflections (the amount depends on the directivity characteristics of the microphone). To have a sufficiently strong direct sound component, the microphone should be at least within the reverberant radius to avoid strong coloration effects due to early reflections [8]. The signal C is played back in the playback room. The playback can be described as a convolution of signal C with the binaural room impulse response (BRIR) of the front loudspeakers (fl) to the artificial head. The ATF in the playback room for the left ear is given by $ATF^{(l)}_{\mathrm{fl},C,d}$, were the subscript d denotes the direct sound. The ATF is optimized with frequency dependent gain factors β_i^2 in each gammatone channel i such that the spectral energy distribution is comparable to the energy distribution in the recording room.

$$ATF^{(l)}_{\mathrm{ref},d,i} \quad - \quad \beta_i^2 \quad \cdot \quad ATF^{(l)}_{\mathrm{fl},C,d,i} \quad \approx \quad 0 \quad (1)$$

In addition two omnidirectional microphones B are placed at two distant positions behind the listener. These microphones record only the diffuse sound field resulting from the sound source. The signal B predominantly contains reflections and only a small amount of direct sound. The distance between the two diffuse field microphones is set to 5 m to avoid high correlations between the signals $B^{(l)}$ and $B^{(r)}$. This enables cross-mixing the two signals to achieve the desired cross correlation. Two dipole loudspeakers, are located in the back of the listener at similar positions as in a 5.1 multichannel setup [10]. These dipoles excite only the diffuse field in the playback room. Thanks to its figure of eight directivity characteristic, the listener only receives lateral reflections from the dipoles and will not perceive the dipoles as separate sound sources. The two signals $B^{(l)}$ and $B^{(r)}$ are convolved with the BRIR measured from the two dipole loudspeakers (h_{dip}) to the artificial head in the playback room. The resulting ATF is $ATF^{(l)}_{\mathrm{dip},B,i}$ for each gammatone-channel i. [2] found an analytical solution how the reverberant sound field has to be excited such that the overall spectral energy distribution in the playback room is comparable to that of the recording room:

$$ATF^{(l)}_{\mathrm{ref},rev,i} - \alpha_i^2 \cdot ATF^{(l)}_{\mathrm{dip},B,i} - \beta_i^2 \cdot ATF^{(l)}_{\mathrm{fl},C,rev,i} \approx 0 \quad (2)$$

Here, $ATF^{(l)}_{\mathrm{ref},rev,i}$ is the reverberant path in the recording room, β_i^2 is a frequency dependent gain factor to control the amount of the diffuse field, and $ATF^{(l)}_{\mathrm{fl},C,rev,i}$ is the reverberant path of the convolution product of signal C and $h^{(l)}_{\mathrm{fl}}$. The separate excitation of the diffuse field gives the advantage that the reverberant path of $ATF^{(l)}_{\mathrm{fl},C,rev,i}$ and with that effectively also the directivity characteristics of the front

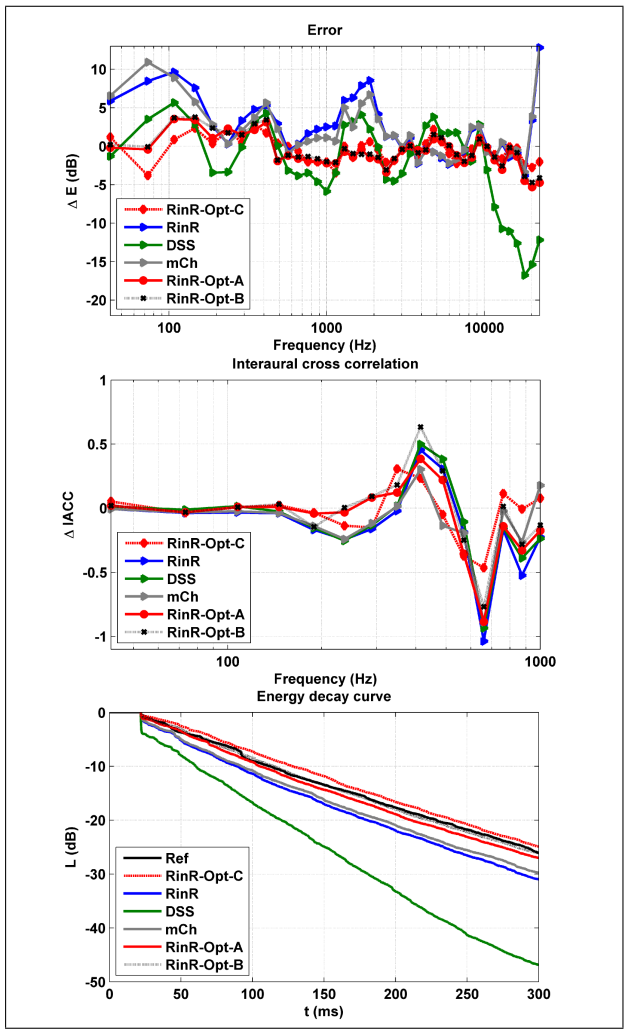

Figure 2. Objective evaluation for all rendering conditions measured at the artificial head in the playback and recording room. Top panel: Spectral energy error for RR1 according to Eq. 5. Middle panel: ΔIACC. An ΔIACC below zero means a higher IACC compared to the reference, above zero means a lower IACC compared to the reference. Bottom panel: Energy Decay Curve (EDC) for all rendering conditions and the reference from the recording room.

loudspeakers is an integral part of the optimization problem. As a result the characteristic of the reproduced sound source is more similar to the source characteristic in the recording room. The interaural cross correlation is adjusted with the inter-channel mixing procedure:

$$B(t)_i^{l'} = B(t)_i^l + \kappa_i \cdot B(t)_i^r \quad (3)$$

$$B(t)_i^{r'} = B(t)_i^r + \kappa_i \cdot B(t)_i^l \quad (4)$$

where κ_i is optimized using a grid search in the range of -1 and 1.

3. Objective Evaluation

This Section describes the objective evaluation of the artificial head signals in terms of the optimized

parameters. Two recording rooms and one playback room were measured with a log-sweep function to obtain the (binaural) room impulse responses which are shown in Fig. 1. Recording room 1 (RR1) was an unoccupied lecture room with a reverberation time of T_{60} = 711 ms, Recording room 2 (RR2) was an unoccupied church with a reverberation time of T_{60} =3040 ms. In the playback room, all binaural room impulse responses were measured between each loudspeaker and the artificial head. The objective evaluation was derived from the BRIRs which were also used in the optimization process. Fig. 2 shows the spectral energy error between Recording room 1 and the playback room for six different rendering conditions. The error is defined as:

$$\Delta E = 10 \cdot log_{10}(\frac{ATF_{\text{ref}}}{ATF_{\text{pr}}}) \tag{5}$$

In this figure, RinR is the conventional playback without optimization, DSS is the dry source signal which is rendered without a recording room influence within the playback room, mCh is the multichannel playback where lateral reflection are excited separately over conventional rear-loudspeakers, Opt-A is the perceptual optimized sound field which was suggested by [2], Opt-B is the perceptual optimized sound field without the interaural cross correlation optimization stage and Opt-C is the perceptual optimized sound field where the dipole loudspeakers were replaced by conventional rear-loudspeakers like those used in the mCh condition. It can be seen that the conditions RinR and mCh have a rather high energy error which is expected to be perceived as an increase in spectral coloration. Considering the error of the proposed method (Opt-A/B/C) it can be seen that the error is fairly small across all frequencies.
Fig. 2 shows the difference in the interaural cross correlation ΔIACC = IACC$_{\text{ref}}$−IACC$_{\text{pr}}$. In sound reproduction it is necessary to accurately reproduce the correct IACC, especially in low frequencies (below 1kHz) when high correlation values occur. Considering the DSS, RinR and mCh condition, it can be seen that ΔIACC is significant different from 0 which will be perceived as a change in the apparent source width. Difference tend to be larger than the optimized conditions. If the condition Opt-A (optimization of all parameters) is compared with Opt-B (without IACC optimization), Opt-B show similar differences. Opt-C has a slightly lower ΔIACC because of the high directivity of the rear-loudspeakers. Fig. 2 shows the energy decay curve of the reference Ref and the various reproduction conditions. The condition DSS, which is basically only the EDC of the playback room, shows the steepest descent and has a reverberation time of T_{60} = 360 ms. RinR has a reverberation time of T_{60} = 592 ms which is a bit lower than the mCH EDC and reverberation time of T_{60} = 617 ms. Opt-A (T_{60} = 720 ms), Opt-B T_{60} = 733 ms and Opt-C

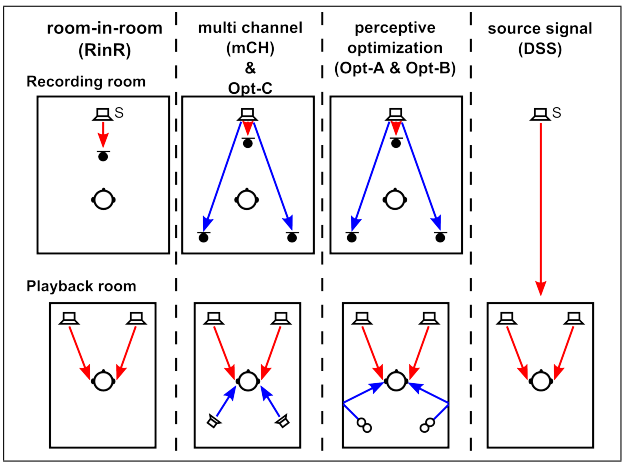

Figure 3. Rendering conditions which are used in the subjective evaluation over loudspeaker. RinR represents a conventional stereo playback, mCH a multi-channel reproduction where lateral reflections are reproduced over rear-loudspeakers, Opt-A, Opt-B are the perceptual reproduced sound field with optimization and without the optimization of the IACC, DSS represents the dry source signal which is rendered directly into the playback room.

(T_{60} = 720 ms) are in good agreement with the reference (T_{60} = 711 ms) and the difference of the reverberation time is smaller than the just noticeable difference in reverberation time of ≈ 20% [9]. The objectives for RR2 show similar results and will not further be discussed.

4. Subjective Evaluation

This Section describes the subjective listening test. A scaling experiment was designed to compare the proposed method with several other rendering conditions over loudspeaker. [2] showed in a headphone experiment that the authenticity, based on comparing the recording room sound field with the sound field resulting from the proposed method, is rather high. The comparison showed the difference to be only just noticeable. In the current listening test, two different recording room RR1 (T_{60}, a small lecture room) and RR2 (T_{60}, a church) were simulated in a playback room. The various rendering conditions which were used in the listening test can be seen in Fig. 3. The room-in-room (RinR) condition simulates a conventional playback without additional excitation of the diffuse field or optimization. The multi-channel (mCh) condition simulates an unprocessed 4.0 reproduction (without center speaker and subwoofer) where the lateral reflections are excited over conventional rear-loudspeakers. The front-channel to back-channel power-ratio was set to 0 dB for RR1 and to 4.5 dB for RR1. These ratios were obtained analyzing (commercial) musical DVD's which were recorded in similar reverberant recording rooms. In addition to our proposed approach (Opt-A), a perceptual optimized sound field is presented without the optimization of

Table I. Bipolar adjective pairs for sound field description used in the listening test.

Adjective		
reverberance	reverberant	dry
coloration	bright	dark
naturalness	natural	unnatural
apparent source width	broad	narrow
listener envelopment	at all	not at all
distance	far away	close
natural coloration	natural	unnatural
spatial naturalness	natural	unnatural

the IACC (Opt-B). To investigate the importance of the directivity of the dipole loudspeakers, in the Opt-C condition, the dipole loudspeakers were replaced by conventional rear-speakers similar as in the mCh-condition. As an anchor signal the completely dry source signal (DSS) was used.

4.1. Attributes

In the subjective evaluation eight different perceptual attributes were used. This contrasts to the previous study which evaluated the resulting authenticity of the proposed optimization method. The list of bipolar adjectives are similar to those which were developed in the SAQI [11] vocabulary. The descriptors were selected in terms of the description of the rendered sound source as well as the description of the reproduced sound field. This allows for have a more detailed view on the perceptual impact of various reproduction methods as well as strengths and weaknesses of these methods. The adjectives can be seen in Table I.

4.2. Stimuli and Subjects

The excerpts used were three different monaural anechoic recordings of musical instruments and speech with a duration of 5-10 seconds. The excerpts were a clean female voice, a picked guitar, and a violin. All stimuli were convolved with the close microphone C and the omnidirectional microphones B. These signals were then played back over the front loudspeakers and the dipole loudspeaker for the condition Opt-A/B. In case of the condition mCh and Opt-C the signals B were played back over the conventional rear-speakers. Eight normal hearing and musical interested subjects with a mean-age of 30 participated in the experiment. The task of the participants was to rate all rendering conditions on a 5-point scale for eight different spatial attributes. The subjects were able to switch in real-time between the rendering conditions in order to allow a direct comparison between the various rendering conditions. The subjects were allowed to listen

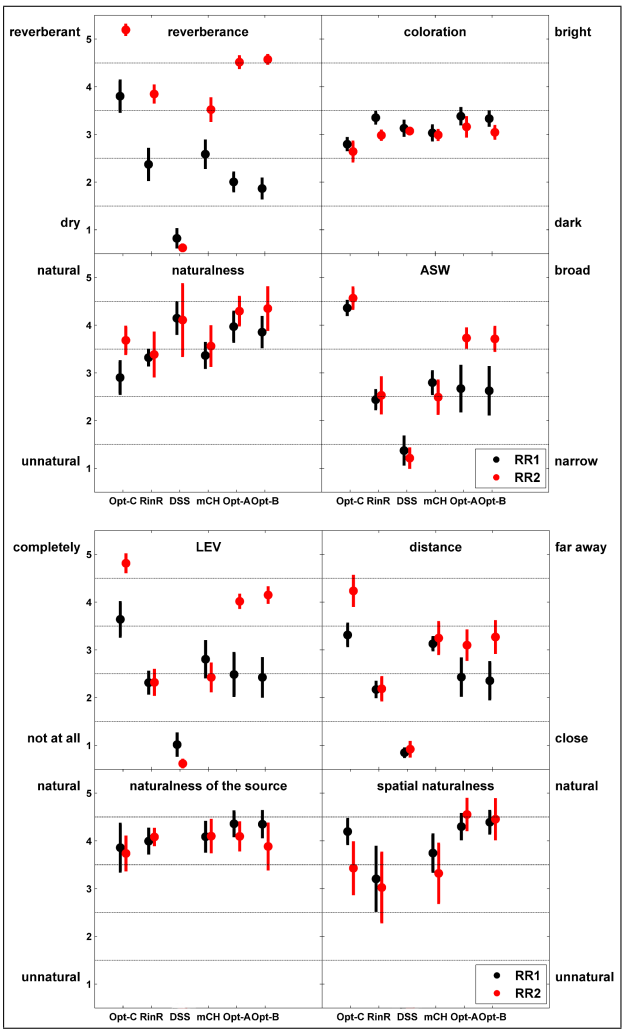

Figure 4. Results of the subjective evaluation. Illustrated are the mean scores for Recording room 1 (RR1) and Recording room 2 (RR2) with standard errors. The rendering conditions are those illustrated in Fig. 3.

as often as they liked to the different rendering conditions presented in the listening test. The detailed positions of all microphone positions as well as the room dimensions are shown in [2].

4.3. Subjective results

Figure 4 shows the results of the Scaling-experiment for the simulation of recording room RR1 (black) and recording room RR2 (red) for the eight attributes. Illustrated are the mean scores with the standard error across seven participants. The dry source signal (DSS) was rated for all attributes, except for the *naturalness of the source* and the *spatial naturalness*. All other conditions were evaluated for all attributes.

Considering the attribute *reverberance* for RR1, it is visible that RinR and mCh are slightly more reverberant as compared to Opt-A and Opt-B. Opt-C was evaluated with a high reverberance and DSS which contains only the room-acoustics of the playback room was evaluated dry. A similar trend can be

seen for RR2. Opt-C was evaluated highly reverberant, similar to Opt-A and Opt-B. RinR and mCh was evaluated a bit lower. Considering the *coloration* it can be seen that there are no significant differences between the rendering conditions for both reproduced rooms. In terms of *naturalness* the conditions DSS and Opt-A/B were rated slightly higher compared to RinR, mCh and Opt-C. A similar trend can be seen for RR2 except for Opt-C. The evaluation of *ASW* for RR1 shows that the DSS condition was evaluated with a narrow source width. RinR, mCH, Opt-A/B were rated with a similar source width, only Opt-C shows a broad source. Considering RR2, the optimized sound conditions show a high tendency to make the source much broader. The evaluation of the *LEV* shows that especially for RR2 and the optimized conditions the subjects feel highly enveloped from the sound source. For RR1 the effect of *LEV* is smaller and similar to the RinR and mCh condition. The perceived *distance* shows an increase between DSS and RinR. The mCh condition are perceived a little more far away compared to Opt-A/B. Opt-C was perceived in a similar distance to the mCH condition. For RR2 Opt-C was perceived at a higher distance, but mCh and Opt-A/B were evaluated similar. Only RinR and DSS were perceived closer. The *naturalness of the source* was evaluated for RR2 as natural across all conditions. For RR1, Opt-A/B show a slightly higher naturalness compared to RinR, mCh or Opt-C. In terms of the *spatial naturalness* Opt-A/B were evaluated highest compared to RinR, mCh or Opt-C. This tendency can be seen for both simulated recording rooms.

5. Discussion

The subjective evaluation over loudspeaker showed, that the optimized sound field improves the perception in terms of the spatial naturalness as well as the *listener envelopment* of the presented scene, and partially also the *apparent source width*. Even if conventional rear-loudspeakers are used, these improvements are still present.

When lateral reflections are added via conventional rear-loudspeakers (mCh), the *spatial naturalness* is rated much lower compared to the optimized reproduction. Since we specifically optimize the IACC, the perceived envelopment is increased compared with the RinR or mCH condition especially for the RR2. The comparison between Opt-A and Opt-B has shown that the optimization of the IACC has only a small perceptual effect. In the technical evaluation of the errors in the IACC it can be seen that the Opt-A and Opt-B conditions differ very little. Probably the energetic optimization is sufficient for obtaining the desired IACC.

The replacement of the dipole loudspeakers with conventional rear-speakers (Opt-C) shows that the *LEV* and *reverberance* ratings are highest but it was rated much lower in the overall *naturalness* and the *spatial naturalness*. A reason for that is the violation of a general assumption. In the design of this approach we assume no direct sound from the dipole loudspeakers because they are aligned in such a way that the listener receives no direct sound. The replacement by conventional rear-loudspeakers has the consequence that the listeners will hear an additional direct sound component which can be perceived as separate sound sources (reported by subjects). This decreases the naturalness because it leads to the possibility to locate the same sound source on three different locations. This detrimental effect on the plausibility of the reproduced scene shows the advantage of exciting the diffuse field with the dipole loudspeakers.

Acknowledgement

The authors would like to thank the Deutsche Forschungsgemeinschaft for supporting this work as part of the Forschergruppe Individualisierte Hoerakustik (FOR-1732)

References

[1] J. Breebaart and S. van de Par and A. Kohlrausch and E. Schuijers: Parametric Coding of Stereo Audio. Journal on Applied Signal Processing 9 (2005), 1305-1322.

[2] J. Grosse, S. van de Par. Perceptually accurate reproduction of recorded sound fields in a reverberant room using spatially distributed loudspeakers. IEEE Journal of Selected Topics in Signal Processing, vol.PP, no.99

[3] J. S. Bradley and G.A. Soulodre: Objective measures of listener envelopment. J. Acoust. Soc. Am. 98 (1995), 2590-2597.

[4] C. C. J. M. Hak and R. H. C. Wenmaekers.: The Impact of Sound Control Room Acoustics on the Perceived Acoustics of a Diffuse Field Recording. Trans. Sig. Proc., WSEAS 2010, 175-185.

[5] M.R. Schroeder: Statistical Parameters of the Frequency Response Curves of Large Room. J. Audio Eng. Soc. 35 (1987), 299-306.

[6] Brian C.J. Moore: An introduction to the psychology of hearing. Brill, 2012.

[7] V. Hohmann: Frequency analysis and synthesis using a Gammatone filterbank. Acta Acustica united with Acustica 88 (2002) 433-442.

[8] A. Haeussler and S. van de Par: Theoretischer und subjektiver Einfluss des Aufnahmeraumes auf den Wiedergaberaum. DAGA 2014, 40. Jahrestagung fuer Akustik, Oldenburg.

[9] Zihou Meng and Fengjie Zhao and Mu He: The Just Noticeable Difference of Noise Length and Reverberation Perception. International Symposium on Communications and Information Technologies, ISCIT 2006, 418-421.

[10] ITU-R Recommendation BS.775-3: Multichannel Stereophonic Sound System with and without Accompanying Picture. International Telecommunication Union, Switzerland, Geneva, 2012.

[11] A. Lindau et al. A Spatial Audio Quality Inventory (SAQI). Acta Acustica united with Acustica 100.5 (2014): 984-994.