PERCEPTUALLY OPTIMIZED ROOM-IN-ROOM SOUND REPRODUCTION WITH SPATIALLY DISTRIBUTED LOUDSPEAKERS

Julian Grosse, Steven van de Par

Cluster of Excellence "Hearing4all"
Acoustics Group
Carl von Ossietzky University Oldenburg, Germany

In sound reproduction it is desired to reproduce a recording of an instrument made in a specific room (e.g. a church or concert hall) in a playback room such that the listener has a plausible and authentic impression of the instrument including the room acoustical properties of the recording room. For this purpose a new method is presented that separately optimizes the direct sound field and recreates a reverberant sound field in the playback room that matches that of the recording room. This approach optimizes monaural cues related to coloration and the interaural cross correlation (IACC), responsible for listener envelopment, in both rooms based on an artificial head placed at the listener's positions. The cues are adjusted using an auditorily motivated gammatone analysis-synthesis filterbank. A MUSHRA listening test revealed that the proposed method is able to recreate the perceived room acoustics of the recording room in an accurate way.

Index Terms— virtual acoustics, perceptual optimization, Room-in-Room reproduction

1. INTRODUCTION

The perception of a recording strongly depends on the room where it was recorded and where it is reproduced. There are many method to reproduce a sound field with multiple loudspeakers, e.g. wave-field-synthesis [1], ambisonics [2] or a stereo playback of a recording. Most of these methods can not easily be implemented in a living room because they require many loudspeakers and do not consider the influence of the living room acoustics except e.g. [3].

The perceived quality of an audio signal by rendering a recorded sound source in a reverberant playback room depends on the reverberation time and the perceived coloration of the sound source [4]. The method that is presented in this study does not aim to reconstruct the exact physical transfer function at the eardrum of a listener but rather perceptually relevant parameters that are defined on the scale of critical bands [5]. The perceptual relevant parameters ILD (interaural level differences), ITD (interaural time differences) and the interaural cross correlation (IACC) can sufficiently describe the spatial perception of a stereo audio signal [6] (on a fixed spatial position). Monaural timbre cues can be described in terms of an excitation pattern. These monaural and binaural cues are derived based on artificial head recordings in the recording and playback room with the aim to match the cues of the playback room to the cues of the recording room with a minimum of four loudspeakers in the playback room.

In our approach the direct and the reverberant sound field are recorded, optimized, and reproduced separately which enables control of the reverberation time, the overall coloration as well as IACC in the playback room relative to the recording room. In a listening test the reference signal of two simulated recording rooms are compared with the applied optimization in two different playback rooms.

2. METHOD

This section describes the setup and the method of the applied optimization on the basis of perceptual cues.

2.1. Analysis of the recording room

The three parameters that are optimized are the coloration, the interaural cross correlation and the estimated reverberation time (T60). The first parameter is the spectral energy distribution across auditory critical bands measured on an artificial head in the recording room. Fig. 1 shows the experimental setup used in this study. With the aim to reproduce the direct and the diffuse path separately, the binaural room impulse response (BRIR) at the "ear-drum" of the artificial head in the recording room $h(t)_{\mathrm{ref}}$ in (1) can be split into a direct and a diffuse path:

$$h(t)^{(\mathrm{l})}_{\mathrm{ref}} = h(t)^{(\mathrm{l})}_{\mathrm{ref,d}} + h(t)^{(\mathrm{l})}_{\mathrm{ref,rev}} \tag{1}$$

The subscript d,rev are related to the direct sound and to the reverberative sound in the recording room, respectively. The derivation for the right ear only requires l to be replaced by r. The separated parts are filtered by a 4^{th}-order gammatone filterbank described in [7]. The bandwidth of the filters follows an ERB-spaced distribution (equally rectangular bandwidth) which is related to critical bands in the human auditory system. The energy of the gammatone filtered impulse response

This work was supported by the DFG Forschergruppe Individualisierte Hoerakustik (FOR-1732).

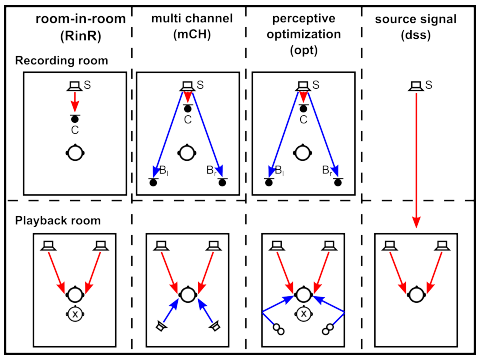

Fig. 1: Experimental setup. RinR: Rendering of the direct sound (red) over the front loudspeakers. mCH: Rendering of the direct sound (red) over the front loudspeakers and the diffuse signals (blue) over rear-loudspeaker. Opt: Rendering of the direct sound (red) over the front loudspeakers and the diffuse signals (blue) over dipole-loudspeakers. Dss: Rendering of the source signal directly in the playback room. The artificial head in the playback room was placed in a 60 ° stereo triangle. The position marked with an X are 0.5 m away from the sweet-spot.

$h(t)$ in the i^{th}-band is defined by (2):

$$\langle|\gamma_i(h(t))|^2\rangle \tag{2}$$

assuming that $h(t)$ is a real-valued signal. The "auditory transfer function" (ATF) of the left ear (l) for the direct sound (d) of the recording room (ref) is determined by (3):

$$\langle|\gamma_i(h(t)^{(\mathrm{l})}_{\mathrm{ref,d}})|^2\rangle = \int_{t_d}^{t_\mathrm{m}} \int_{-\infty}^{+\infty} |h(\tau)^{(\mathrm{l})}_{\mathrm{ref}} * \gamma_i(t-\tau)|^2 d\tau \tag{3}$$

where t_d is the start time of the direct sound. The resulting ATF in (3) contains the energy in each frequency band i. The separation time constant t_m determines the direct-to-diffuse ratio and therefore the reverberation time T_{60} in the playback room. If a relatively small t_m is chosen the energy of the direct sound is low compared to the energy of the reverberative part and will grow if the separation time constant increases. The ATF of the reverberant part is determined in a similar way by integrating the diffuse part of the BRIR in the recording room from t_m to the end of the impulse response.

The second parameter is the IACC and in this context the normalized cross correlation of the time-signal of the whole gammatone filtered BRIR $\gamma_i(h(t))$ in the recording room is regarded.

The third parameter is the estimated T_{60} reverberation time which is determined by interpolating the energy decay curve (edc) from -5 to -35 dB. To determine the T_{60} in the recording room, mean T_{60} of the left and the right artificial head signal is observed.

2.2. Analysis of the playback room

The analysis in the playback room is similar to the analysis in the recording room.

The BRIR $h(t)^{(\mathrm{l})}_{\mathrm{play}}$ can be divided again with the same separation time constant t_m into a direct and a diffuse path ($h(t)^{(\mathrm{l})}_{\mathrm{play,d}}$,$h(t)^{(\mathrm{l})}_{\mathrm{play,rev}}$). The microphone $C(t)$, which records the direct sound close to the sound source, is convolved with the BRIR of the front loudspeaker to the artificial head ($C(t)*h(t)_{\mathrm{play}}$). The diffuse microphone signals $B(t)^{(\mathrm{l})}$ and $B(t)^{(\mathrm{r})}$, record the sound source at two distant positions in the recording room, are convolved with the dipole loudspeaker signals $h(t)^{(\mathrm{l})}_{\mathrm{dip}}$ and $h(t)^{(\mathrm{r})}_{\mathrm{dip}}$. Due to the directivity pattern of the dipole loudspeakers, the reverberant field is independently excited when the zero is directed towards the listener. The BRIR $h(t)^{(\mathrm{l})}_{\mathrm{pr}}$ consisting of the direct and diffuse sound in the playback room can be written as:

$$h(t)^{(\mathrm{l})}_{\mathrm{pr}} = \beta^{(\mathrm{l})} \overbrace{\left[C(t) * h(t)^{(\mathrm{l})}_{\mathrm{play,d}} + C(t) * h(t)^{(\mathrm{l})}_{\mathrm{play,rev}}\right]}^{C(t)*h(t)_{\mathrm{play}}} + \alpha^{(\mathrm{l})} \left[(B(t)^{(\mathrm{l})} * h(t)^{(\mathrm{ll})}_{\mathrm{dip}}) + (B(t)^{(\mathrm{r})} * h(t)^{(\mathrm{rl})}_{\mathrm{dip}})\right] \tag{4}$$

where $h(t)^{(\mathrm{l})}_{\mathrm{play},d}$ is the direct sound path and $h(t)^{(l)}_{\mathrm{play},rev}$ the diffuse sound path of $h^{(\mathrm{l})}_{\mathrm{play}}$ to the left ear. The weighting factor β is introduced to adjust the direct sound and α allows to adjust the overall energy in the playback room. This optimization based on the auditory transfer function is perceptually inspired and does not aim to optimize the real physical sound signal at the "listener's-eardrum". Therefore the aim is to approximate (5):

$$\frac{\langle|\gamma_i(h(t)^{(\mathrm{l})}_{\mathrm{ref}})|^2\rangle}{\langle|\gamma_i(h(t)^{(\mathrm{l})}_{\mathrm{pr}})|^2\rangle} = \Delta E \approx 1 \tag{5}$$

By combining (4) of the playback room and (1) of the recording room with (5) the outcome for the direct path in the playback room is:

$$\beta^2_{i,\mathrm{l}} \cdot EP^{\mathrm{play,l}}_{\mathrm{C,d},i} = EP^{\mathrm{ref,l}}_{\mathrm{d},i} \tag{6}$$

where:

$$EP^{\mathrm{play,l}}_{\mathrm{C,d},i} = \langle|\gamma_i(\overbrace{C(t) * (h(t)^{(\mathrm{ll})}_{\mathrm{d,play}} + h(t)^{(\mathrm{rl})}_{\mathrm{d,play}})}^{d(t)})|^2\rangle$$
$$EP^{\mathrm{ref},l}_{d,i} = \langle|\gamma_i(h(t)^{(l)}_{d,\mathrm{ref}})|^2\rangle$$

The superscript ll refers to the path from the left loudspeaker to the left "ear" and rl from the right loudspeaker to the left "ear". The outcome for the reverberant path is:

$$\alpha^2_{i,\mathrm{l}} \cdot EP^{\mathrm{play,l}}_{\mathrm{dip,rev},i} = EP^{\mathrm{ref,l}}_{\mathrm{rev},i} - \beta^2_{i,\mathrm{l}} \cdot EP^{\mathrm{play,l}}_{\mathrm{C,rev},i} \tag{7}$$

where:

$$EP^{\mathrm{ref,l}}_{\mathrm{rev},i} = \langle|\gamma_i(h(t)^{(\mathrm{l})}_{\mathrm{ref,rev}})|^2\rangle$$
$$EP^{\mathrm{play,l}}_{\mathrm{C,rev},i} = \langle|\gamma_i(C(t) * (h(t)^{(\mathrm{ll})}_{\mathrm{play,rev}} + h(t)^{(\mathrm{rl})}_{\mathrm{play,rev}}))|^2\rangle$$
$$EP^{\mathrm{play,l}}_{\mathrm{dip,rev},i} = \langle|\gamma_i((B(t)^{(\mathrm{l})} * h(t)^{(\mathrm{ll})}_{\mathrm{dip}}) + (B(t)^{(\mathrm{r})} * h(t)^{(\mathrm{rl})}_{\mathrm{dip}}))|^2\rangle$$

In order to solve the values for the scaling factors α and β one has to consider that gammatone filters are overlapping. As a consequence the simple approach of using the energy ratios can not be applied. In order to solve (6) and (7) the energy of the neighbouring filters has to be considered.

2.3. Filter coefficient processing

For solving the coefficients α and β a method was proposed by van de Par et al. [8]. This method is controlled by energy based ATF's and provides real-valued solutions. This solution was used because of fast convergation of the algorithm. Equation (8) defines a spreading matrix:

$$\Gamma_{i,j} = \sum_{f} |\gamma_i|^2 | \sum_{j=1}^{N} \gamma_j(f) D(f)_{\text{play}}|^2 \tag{8}$$

$\Gamma_{i,j}$ is the energy of each filter-band i with the energy of the neighbouring sub-bands j. The sum over all sub-bands j of $\Gamma_{i,j}$ is the overall energy in filter i of N-bands. The spreading matrix represents how a spectral band is spread across the gammatone filters. $D(f)_{\text{play}}$ is the frequency representation of $d(t)_{\text{play}}$ which corresponds to the transfer function of the direct sound of the front loudspeakers in the playback room. $\gamma_i(f)$ and $\gamma_j(f)$ are the transfer function of the gammatone filterbank. Secondly we define a backward spreading matrix $H_{j,i}$ in (9). $H_{j,i}$ is a transposed version of the spreading matrix and normalized to the total energy over the sub-bands j:

$$H_{j,i} = \frac{\Gamma_{i,j}}{\sum_{j=1}^{N} \Gamma_{i,j}} \tag{9}$$

The optimal filter coefficients β and α are obtained with algorithm 1.

Algorithm 1: Calculation of the optimal coefficients

Input: $\Gamma_{i,j}$, $EP_{d,i}^{\text{ref,l}}$ or $EP_{\text{rev},i}^{\text{ref,l}}$
Output: β_j^2
Initialisation: $\beta_j^2 = 1$ for j=1,...,N
while $k \leq K$ *(e.g.* $K = 20$*)* **do**

$$\hat{E}_i = \sum_{j=1}^{N} \Gamma_{i,j} \beta_j^2 \tag{10}$$

$$\varepsilon_i = \frac{EP_{d,i}^{\text{ref,l}}}{\hat{E}_i} \tag{11}$$

$$c_j = \sum_{i=1}^{N} H_{j,i} \varepsilon_i \tag{12}$$

$$\beta_{j,k}^2 = \beta_j^2 c_j \tag{13}$$

$$k \leftarrow k + 1 \tag{14}$$

end
return $\beta_{j,K}^2$;

2.4. IACC optimization

To recreate the listener envelopment in the playback room the interaural cross correlation must be optimized. Because the IACC depends on the energy optimization of the direct and the diffuse path, firstly the energy in the direct path must be adjusted and finally the IACC over the dipole loudspeakers will be adapted. For this reason the IACC is processed iteratively by mixing the recorded diffuse field signals B_i^l and B_i^r in the following way:

$$B_i^{\text{l}'} = B_i^{\text{l}} + \kappa_i \cdot B_i^{\text{r}} \text{ and } B_i^{\text{r}'} = B_i^{\text{r}} + \kappa_i \cdot B_i^{\text{l}} \tag{15}$$

where κ_i is the mixing coefficient which is iteratively varied with a stepsize of 0.2 in a range from [-1:1]. The mixed signals are $B_i^{\text{l}'}$ and $B_i^{\text{r}'}$. Because the ATF of the dipole loudspeakers depend on the mixing coefficient κ, the processing has to be solved with algorithm 2.

Algorithm 2: Optimization of the IACC

for $\kappa_i = -1$ ***to*** 1 *with a stepsize of 0.2* **do**

1. Solve (6) to adjust the direct sound that the energy is comparable to the direct sound in the recording room.
2. Mixing the diffuse field microphone signals B according to (15) with κ_i.
3. Process the energy of the dipole loudspeakers that the overall energy is comparable to the energy in the recording room.
4. Process the IACC of the playback room.

end
return κ_i
The best suitable κ_i is chosen in this grid search which minimizes the IACC difference between the recording and the playback room
(arg min[$\text{IACC}_i^{\text{rec}} - \text{IACC}(\kappa_i)_i^{\text{play}}$] ≈ 0).

2.5. Synthesis: Loudspeaker & Headphone

The synthesis stage introduced in [7] was used. The square-root of the processed coefficients β_i and α_i are multiplied in the time or frequency domain as a real-valued gain factor on each channel i. The optimized signals $C(t)_{\text{opt}}$ and $B(t)_{\text{opt}}$ are rendered in the playback room and in case of the listening test the optimized signals are convolved with the particular BRIR of the playback room.

3. OBJECTIVE EVALUATION

In this Section the optimization algorithm as well as the optimized parameters are discussed.

3.1. Evaluation: Algorithm

In Fig. 2 the ratio ε_i according to (11) as a function of iterations is illustrated. Considering the ATF $\varepsilon_{i,20}$ it can be seen that it adapts rather well to zero and the error is fairly small. Fig. 2 illustrates the normalized-root-mean-square error (nRMSE) of the estimated ATF relative to the target ATF, which is defined:

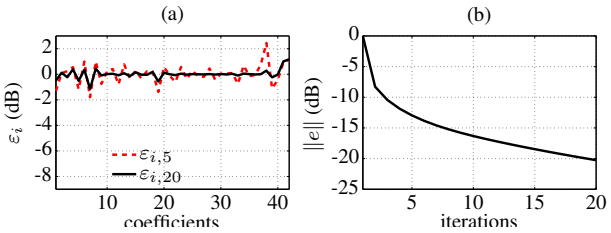

Fig. 2: (a): $\varepsilon_{i,5}$ (red,dotted) and $\varepsilon_{i,20}$ (black,solid) shows the ratio according to (11) after 5 respectively 20 iterations. (b): nRMSE as a function of iterations according to (16).

$$\|e\| = \frac{\sqrt{\sum_i (EP_{\mathrm{d},i}^{\mathrm{ref,l}} - \hat{E}_i)^2}}{\sqrt{\sum_i (EP_{\mathrm{d},i}^{\mathrm{ref,l}})^2}} \tag{16}$$

By comparing the estimated ATF's as a function of the iterations, it can be seen that the solution converges in the first few iterations relatively fast and will change only slightly by increasing the number of iterations. By increasing the number of iterations, the RMSE $\|e\|$ does not always converge to the global minimum. This can be explained by the lack of statistical independence between the filters. Because of the overlapping filters, there is a high correlation between neighbouring filters.

3.2. Evaluation: Optimized parameters

Fig. 3 shows the optimized parameters of the lecture room simulation in the loudspeaker laboratory. It is visible (c.f. Fig. 3 (a)) that the spectral error (ΔE according to (5)) of our approach (Opt) is fairly small. The simple room-in-room method has a higher spectral error distribution which can be perceived as an increase in coloration compared to the reference signal in the recording room. The IACC of the room-in-room (RinR) method is, in comparison to the reference IACC in the recording room, significantly different. As it can be recognized is our approach agrees much more closely with the reference IACC than the RinR method does. Fig. 3 (c) illustrates the estimated reverberation time T_{60} of the recording room, the applied optimization (Opt) and RinR method. It shows that the RinR method has a lower reverberation time ($T_{60} = 597$ ms) in contrast to the applied optimization and that our approach matches ($T_{60} = 706$ ms) well to the reverberation time of the recording room ($T_{60} = 699$ ms).

4. SUBJECTIVE EVALUATION

A MUSHRA test is used to evaluate several rendering methods which are shown in Fig. 1 relative to the proposed method (noted as Opt). For this two recording rooms (Lecture room ($T_{60} = 699$ ms) and a Church ($T_{60} = 3040$ ms)) and two playback rooms (PBR 1 ($T_{60} = 371$ ms) and PBR 2 ($T_{60} = 697$ ms)) are used. The room-in-room condition simulates a conventional stereo reproduction [9] without optimized lateral reflections whereas the multi-channel condition simulates an unprocessed 4.0 multichannel reproduction (without a center speaker) with lateral reflections over rear loudspeakers. In

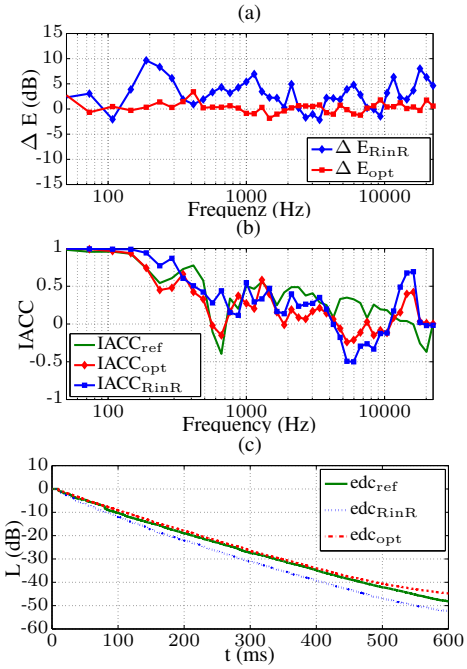

Fig. 3: Example of the optimized parameters for the simulated lecture room in the loudspeaker lab. a): Shows the spectral energy distribution of the recording room and the playback room. ΔE_{opt} illustrates the error of the perceptive approach, ΔE_{RinR} illustrates the error of a conventional room-in-room stereo reproduction. b): Shows the IACC of the recording room (ref), the perceptive approach (opt) and the room-in-room (RinR) reproduction. c): Shows the energy decay curve (edc) of the recording room, the RinR reproduction and perceptual optimization.

this condition the front-to-back ratio was derived from musical DVD's; a ratio of 0 dB was applied for the simulated lecture room and 4.5 dB for the simulated church. The first anchor signal is the 3.5 kHz low-pass filtered reference signal as described in the ITU-R Recommendation BS.1534-1. In addition to the low-pass filtered anchor, a spatial anchor was used that represents the dry source signal in the playback room. To investigate the spatial robustness of our approach the artificial head was moved 0.5 m back with respect to the front loudspeaker in order to investigate a position away from the sweet spot (noted as $\mathrm{Opt_x}$ and $\mathrm{RinR_x}$).

4.1. Stimuli and subjects

The listening test was performed by 12 normal hearing subjects (2 female, 10 male) with a mean age of 29 years. The excerpts used to derive the stimuli were twelve 5 to 10 seconds long, anechoic mono signals which covered a broad range of instruments. These signals were convolved with the specific BRIR in the playback room. All stimuli were presented at 67 dB-SPL. The task of the subject was to rate the difference of the test conditions on a scale between 0 (large difference) and 100 (no difference) relative to the reference signal. All

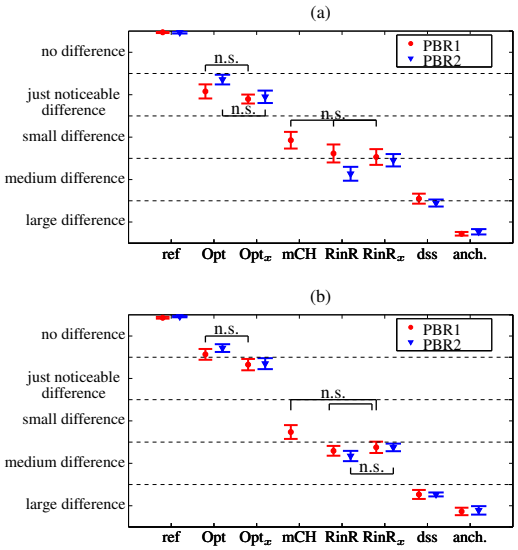

Fig. 4: Ratings of the subjective evalation in PBR 1 (circles) and PBR 2 (triangles). Illustrated is the simulated lecture room (a) and the simulated church (b). The symbols represent the mean-scores and the error bars show the standard error over the the different types of instruments. The marked errorbars are not significant different (n.s.) to $\alpha < 0.05$ compared to the other conditions.

subjects got instructions and had to pass a training phase with several instruments and including all processing conditions. The listening test was done from all subjects for all conditions in four sessions.

4.2. Subjective results

The results of the listening test are shown in Fig. 4. The condition mCH was only rated for one playback room. All other conditions were rated for both playback rooms and recording rooms. The results show that our applied optimization (Opt) was always rated with a smaller difference compared to the reference condition (ref). A view on the condition RinR shows that this rendering method was rated with a relatively large difference for both recording rooms. This difference can be explained by the much lower reverberation time in the playback rooms, higher energy variations over all frequencies and a much higher IACC compared to the recording room. Even in the multichannel condition (mCH), where the diffuse field is excited separately with conventional rear-speakers, the difference was rated much larger as the applied optimization. A comparison of the positions out of the sweet spot of the RinR and the Opt conditions shows that the differences are rather small among these positions and that the proposed method shows a good spatial robustness (large sweet spot). The comparison of the results of different playback rooms show that our method is perceived only with a relatively small difference compared to the same reference signal in the recording room. None significant conditions are shown Fig. 4 as "n.s."

5. CONCLUSION

This study presents a method for rendering a sound source including the room acoustical properties of the recording room in an reverberant playback room. Rather than aiming to optimize the physical sound field in an accurate way this method optimizes a set of perceptually relevant parameters. These parameters, the overall timbre and the IACC, are optimized to perceptually recreate the sound field with a small set of loudspeakers. Because of the placement of an artificial head in both rooms and the specific microphone and loudspeaker arrangement, the direct and reverberative path can be analyzed and adjusted separately. The direct sound will be rendered over a set of stereo-loudspeakers. This enables control of the direction of arrival and the amount of energy that it corresponds to the direct sound of the recording room. Although this was not investigated in this study, in principle the direct sound field could be presented using a wave-field-synthesis system or an ambisoncis system. The reverberative field is excited using the two dipole loudspeakers which allows the control of the overall timbre and the IACC. A subjective comparison showed, that our proposed method was always rated with a higher preference even if the artificial head was moved out of the sweet spot.

REFERENCES

[1] A.J. Berkhout, "A holographic approach to acoustic control," *J. Audio Eng. Soc.*, vol. 36, pp. 977–995, 1988.

[2] M.A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21, pp. 2–10, 1973.

[3] S. Spors, H. Buchner, and R. Rabenstien, "A novel approach to active listening room compensation for wave field synthesis using wave-domain adaptive filtering," *ICASSP*, vol. 4, pp. iv–29– iv–32, 2004.

[4] M.R. Schroeder, "Statistical parameters of the frequency response curves of large room," *J. Audio Eng. Soc.*, vol. 35, pp. 299–306, 1987.

[5] Brian C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th edition edition, 2 January 2012.

[6] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric coding of stereo audio," *Journal on Applied Signal Processing*, vol. 9, pp. 1305–1322, 2005.

[7] V. Hohmann, "Frequency analysis and synthesis using a gammatone filterbank," *Acta Acustica united with Acustica*, vol. 88, pp. 433–442, 2002.

[8] S. van de Par, V. Kot, and N.H. van Schijndel, "Scalable noise coder for parametric sound coding," *118th Convention of the Audio Engineering Society, J. Audio Eng. Soc.*, vol. 53, no. 6465, 2005.

[9] 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," *WSEAS Trans. Sig. Proc.*, vol. 6, no. 4, pp. 175–185, Oct. 2010.