# Evaluation of binaural sound reproduction systems with focus on perceptual quality

Florian Pausch[1], Lukas Aspöck[2], Janina Fels[1,3]

[1] *Institute of Technical Acoustics, Medical Acoustics Group, 52074 Aachen, Germany, Email: {fpa, jfe}@akustik.rwth-aachen.de*

[2] *Institute of Technical Acoustics, RWTH Aachen University, 52074 Aachen, Germany, Email: las@akustik.rwth-aachen.de*

[3] *Forschungszentrum Jülich GmbH (INM-1), 52428 Jülich, Germany, Email: j.fels@fz-juelich.de*

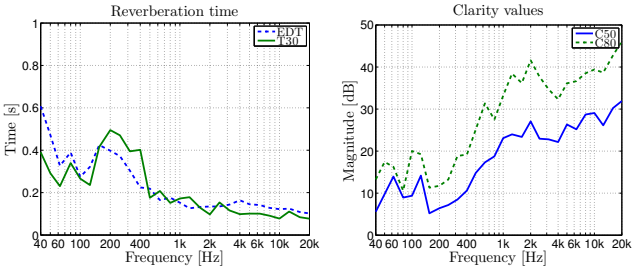
## Introduction

State-of-the-art acoustic virtual reality systems offer various methods to spatialize sound and create complex environments [1, 2, 3]. A disadvantage of several methods is the high amount of hardware which constricts their application range to rooms with generous space. In listening environments with limited space, like hearing booths, methods which require a small loudspeaker setup to synthesize a surrounding sound field are preferred.
Such methods include binaural audio playback through headphones or through loudspeakers in combination with a *Crosstalk Cancellation* (CTC) filter network [4]. Binaural signals are based on *Head-Related Transfer Functions* (HRTFs) using their inherent filter characteristics which enable the localization of a sound source [5]. As intrusive hardware should be avoided in virtual reality systems, a loudspeaker-based binaural playback seems preferable. In combination with a head-tracking system, this enables a more realistic experience and, in case of a multi-user scenario, favors an easier interaction between users [6].
But still, the question of how large the perceived difference between a headphone-based and a loudspeaker-based binaural reproduction arises and has to be investigated experimentally. This paper evaluates these two different binaural audio reproduction systems in a dynamic listening experiment with static and moving stimuli and allowed user interaction. The user has to directly compare the perceptual quality of the two systems by analyzing a presented stimulus with respect to different perceptual attributes to subsequently rate these attributes. The stimulus presentation through headphones is either based on HRTFs, rendering a sound source in free-field conditions, or on *Binaural Room Impulse Responses* (BRIRs) which contain room acoustical information of the reproduction room. In this context, it is also interesting if we can achieve an identical perceptual rating through the integration of BRIRs when presenting the stimuli via headphones. On the other hand, a difference is expected when comparing the HRTF-based headphone reproduction to the loudspeaker-based reproduction.

## Listening environment

The listening experiment is carried out in an acoustically optimized listening booth, at the Institute of Technical Acoustics (ITA) Aachen, with a room volume of $V \approx 10.5\,\mathrm{m}^3$. The room impulse responses are measured at twelve different microphone positions and two source positions using a three-way omnidirectional sound source [7].

The early decay time (EDT) as well as the reverberation time ($T_{30}$) are plotted in Figure 1(a). An increased reverberation time between 150-400 Hz due to room modes can be observed, resulting in a low-frequency coloration which affects the loudspeaker-based binaural reproduction. Regarding the clarity values, which are plotted in Figure 1(b), an excellent speech intelligibility and a high musical transparency can be expected.




**(a)** EDT / $T_{30}$ **(b)** $C_{50}$ / $C_{80}$

**Figure 1:** Measured room acoustics of the listening booth: (a) reverberation time, (b) clarity values.

## System overview

In Figure 2, a flowchart diagram of the two binaural reproduction systems is given. The upper path illustrates the two variants for a headphone-based binaural playback. The virtual sound source is generated by convolving a mono audio file with a selected pair of *setup*-HRTFs or *setup*-BRIRs (nearest neighbor interpolation method). The setup-HRTF set consists of ITA dummy head HRTF measurements with a 3° resolution in azimuth and elevation measured at a radius of 1.86 m. For the setup-BRIR set, the transfer paths between the reproduction loudspeakers and the dummy head microphones are measured. The finally used setup-BRIRs are a combination of the setup-HRTF set for the direct sound, and the measured BRIRs using only the part which covers the early reflections and the late reverberations to encode the room information. To minimize the influence of the headphones, a global headphone equalization filter (Hph-EQ), based on the measurement of individual headphone transfer functions of all participants in the listening experiments, is applied [8].
In the lower part of Figure 2, the loudspeaker-based binaural reproduction is shown. To allow a full user rotation and to achieve an a-priori well-conditioned system matrix [9], also called *playback*-HRTF set, the loudspeakers are arranged as shown at a height of 1.48 m (measured from the center between the drivers to the floor) with an

elevation angle of −15°. A free-field loudspeaker equalization filter (LS-EQ) is used to obtain a flat transducer frequency response.
In order to allow for user interaction, the tracker information of an electro-magnetic head tracker (6 DoF, 120 Hz update rate) is integrated to select the appropriate filter sets which results in an update of setup-HRTFs/BRIRs, the CTC filters and the playback-HRTFs.
As auralization software, *Virtual Acoustics*, a real-time auralization environment with MATLAB interface, developed at the ITA Aachen is used.

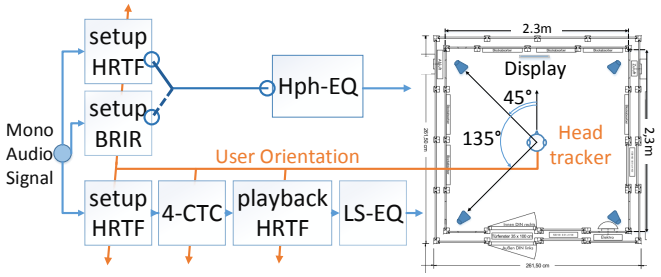



**Figure 2:** Flowchart diagram of the complete system. The upper path describes the two variants of the binaural headphone reproduction (setup-HRTF-/BRIR-based) with headphone equalization (Hph-EQ). The lower path describes the loudspeaker-based binaural reproduction (4-CTC) including a free-field loudspeaker equalization (LS-EQ).

## Crosstalk Cancellation system

The reduction of the acoustic crosstalk paths in the playback-HRTF matrix is realized by a $N$-CTC approach with $N = 4$ loudspeakers all playing simultaneously [4]. The 4-CTC system can be defined compactly using matrix notation, neglecting the loudspeaker transfer functions, through

$$\mathbf{e} = \mathbf{HCb}, \tag{1}$$

with the ear signals $\mathbf{e} = [E_L(z), E_R(z)]^{\mathrm{T}}$, with $(\cdot)^{\mathrm{T}}$ symbolizing the transpose, the playback-HRTF matrix $\mathbf{H}$, the CTC matrix $\mathbf{C}$ and the binaural input signal $\mathbf{b} = [B_L(z), B_R(z)]^{\mathrm{T}}$. All signals are given in complex frequency domain with $z = e^{-j\omega}$. In case of a poorly conditioned matrix $\mathbf{H}$, and to avoid large numerical values, a regularized inverse optimal in a least-square sense is used, which yields

$$\mathbf{C} = \mathbf{H}^{-1} = \mathbf{H}^{\mathrm{H}} \underbrace{(\mathbf{H}\mathbf{H}^{\mathrm{H}} + \beta\mathbf{I})^{-1}}_{\mathbf{Y}} \cdot e^{-z\Delta}, \tag{2}$$

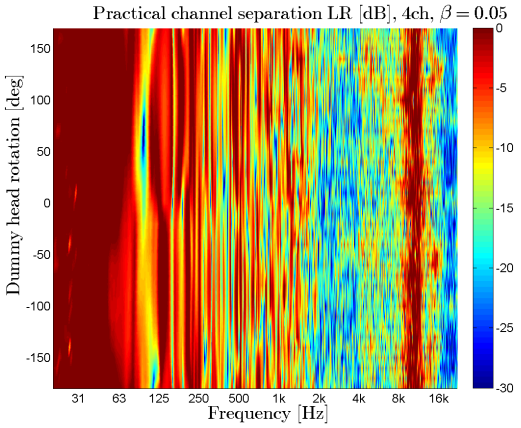
where $(\cdot)^{\mathrm{H}}$ symbolizes the Hermitian of a matrix and $\beta$ is a constant regularization parameter (e.g. $\beta = .05$). The time delay $\Delta$ is necessary to obtain causal CTC filters [4].

### Channel separation

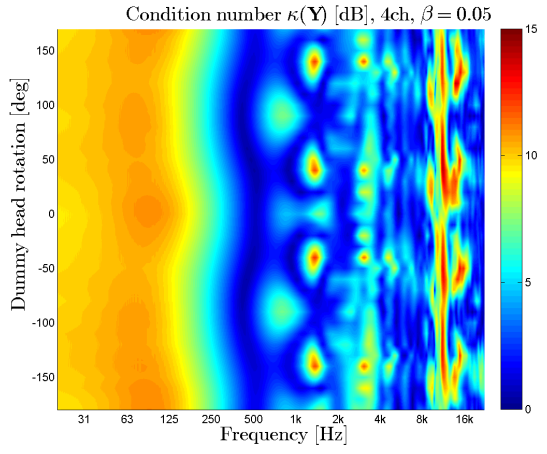
The performance of a CTC system is evaluated through the *channel separation*. Exemplarily, the left-to-right channel separation can be calculated if we set the binaural input signal to $\mathbf{b} = [1, 0]^{\mathrm{T}}$ which simplifies Equation 1 to

$$\begin{aligned} E_L(z) &= \textstyle\sum_{n=1}^{N} H_{nL}(z) C_{Ln}(z), \\ E_R(z) &= \textstyle\sum_{n=1}^{N} H_{nR}(z) C_{Ln}(z), \end{aligned} \tag{3}$$

with $N = 4$. The channel separation is then expressed by the ratio $E_L(z)/E_R(z)$, where negative dB values reflect a higher channel separation.
To get a more complete result, the rotation-dependent channel separation is investigated. This is done by placing a dummy head on a turntable to measure the BRIRs in azimuth steps of 10° using the loudspeakers. Those BRIRs are substituted into Equation 3 as "playback-HRTFs", while employing the original playback-HRTFs for the CTC filter calculation, to obtain the practically achieved channel separation including the room reflections of the listening booth. The results are plotted in Figure 3(a). According to [10], the achieved channel separation lies in a critical range to satisfactorily reproduce the binaural cues.




**(a)** Rotation-dependent left-to-right channel separation.




**(b)** Rotation-dependent condition number.

**Figure 3:** Performance metrics of the 4-CTC system.

### Condition number

As a second performance metric, the *condition number* $\kappa(\mathbf{Y})$ of the matrix $\mathbf{Y}$ in Equation 2 is evaluated. In Figure 3(b), the rotation-dependent condition number $\kappa(\mathbf{Y})$ is shown for all dummy head rotations. The system is well-conditioned for frequencies above 250 Hz with some ill-conditioned spots around 1.5 kHz, 3 kHz and 10 kHz.

## Listening experiment

### Subjects and stimuli

In total, 7 female and 15 male subjects, 21 subjects with and 1 subject without HRTF experience, at the age between 24 and 38 (mean±$\sigma$: 30.2±3.6) participated in the listening experiment.
The stimuli are based on two anechoic mono recordings (music, speech) in wave format, with a sampling rate of 44.1 kHz at a resolution of 16 Bit and a duration of 5.56 s. The virtual source is either *moving* (trajectory 1 and 2) or *static* (trajectory 3) as shown in Figure 4.

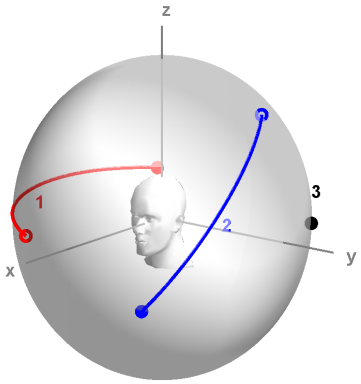



**Figure 4:** Trajectories of the stimuli presented in the listening experiment: Trajectory 1 and 2 cover big circle paths with a length and at a radius of 2 m, where the circle and the dot represent the start and the end of the trajectory, respectively. This results in a source velocity of 0.36 m/s. Trajectory 3 is static.

### Experimental design

The task of the participants is to perceptually rate the two binaural reproduction systems, depicted in Figure 2, in a direct AB-comparison. This is either accomplished by comparing the setup-HRTF-based or the setup-BRIR-based binaural headphone reproduction versus the loudspeaker-based binaural reproduction (comparison type). The parameter set to be rated is a subset from the *Spatial Audio Quality Inventory* (SAQI) [11].
At the beginning of the experiment, all subjects are given an instruction sheet describing the course of events and a second sheet with all involved SAQI parameters including their circumscription. The language is either English or German.
The whole experiment takes about 45 minutes and consists of 12 trials (2 comparison types×2 stimuli×3 trajectories), where each trial contains the rating of 5 SAQI parameters (*difference, externalization, localizability, naturalness, degree-of-liking*), which results in 60 comparisons for each user. A full factorial design with an additional permutation is used to cover all possible combinations and to obtain a unique permutation pool for each user. All subjects are instructed to put the headphones on or off dependent on the selected reproduction type. To include the individual user geometries, the positions of the ear entrances relative to the head tracking sensor are calibrated in advance. The 4-CTC system is thus capable to reproduce the binaural signals exactly at the users' ear entrances.

### Procedure

In Figure 5, the course of events is shown. In an initial training session, the subjects are familiarized with the SAQI parameters and have to pass two trials covering two different stimuli and two trajectories to rate the two comparison types.
After a short break of two minutes, the first part of the experiment starts with the first 6 trials. Thereto, a unique combination (comparison type, stimulus, trajectory) is selected. For stimulus playback, two buttons (A or B) are available, each dedicated to one binaural playback system (headphones or 4-CTC). Subjects are forced to listen to each playback system at least once before being able to rate a SAQI parameter (the maximum number of playbacks is limited to 2). First, the user has to rate the overall difference between the two reproduction systems using an unipolar slider (range: 0 to 3, continuous). If a difference is perceived the rating of the other SAQI parameters will follow using bipolar sliders (range: -3 to 3, continuous). After half of all trials have been passed, another short break of two minutes follows and the second half of the combination pool is queried.

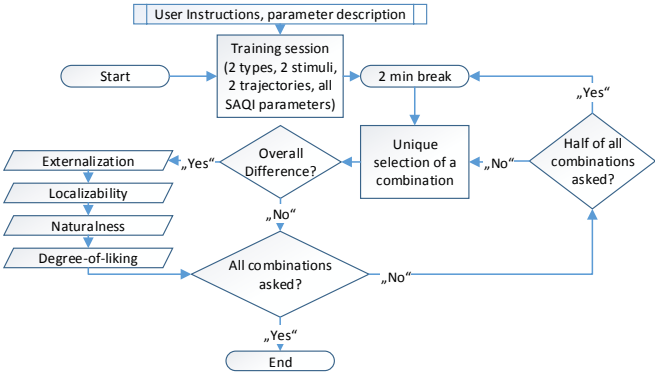



**Figure 5:** Course of events in the listening experiment.

## Results

The results of the listening experiment for different stimulus types including all trajectories for all queried parameters are plotted in Figure 6.
Two different hypotheses are tested in the listening experiment: Hypothesis I assumes that the median difference of the comparison involving the setup-HRTF-based binaural headphone reproduction versus the 4-CTC reproduction ($\Delta_1$) is bigger than the median difference involving the setup-BRIR-based headphone reproduction versus the 4-CTC reproduction ($\Delta_2$), i.e. $\Delta_1 > \Delta_2$. Hypothesis II assumes that the median difference of the comparison involving the setup-BRIR-based binaural headphone reproduction versus the 4-CTC reproduction is zero, i.e. $\Delta_2 = 0$.
A *Wilcoxon rank sum test* revealed that there are no significant differences in median differences for all possible combinations testing Hypothesis I ($\alpha = .05$).
A *two-sided sign test* was applied to test Hypothesis II

($\alpha = .2$). Significant results are found for *externalization* (speech/music, $p = .0869/p = .0065$) and for *localizability* (speech/music, $p = .0000/p = .0000$). The median differences, however, are not significant for *naturalness* (speech/music) and for *degree-of-liking* (speech/music). Additionally, some tendencies in median differences can be observed: The degree of *externalization* is perceived slightly higher in case of a 4-CTC binaural playback. Concerning *localizability*, the 4-CTC reproduction is rated worse in comparison to both headphone-based binaural playback variants, although there are some outliers stating the opposite. This parameter, however, may be improved in the 4-CTC reproduction by selecting a smaller regularization value $\beta$ in Equation 2. The parameters *naturalness* and *degree-of-liking* are rated in preference of the 4-CTC reproduction; this preference is (not significantly) rated higher in comparison to the setup-HRTF-based binaural headphone reproduction than to the setup-BRIR-based binaural headphone reproduction.

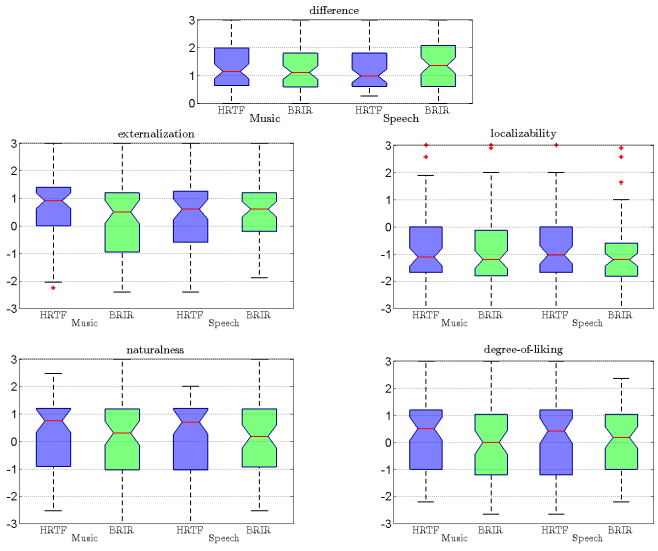



**Figure 6:** Results of the listening experiment: Boxes range from $25^{\text{th}}$ to $75^{\text{th}}$ percentile, whiskers cover $\pm 2.7\sigma$ and notches mark the 95% confidence interval. The slider values are plotted on the y-axis against the different comparison type, e.g. setup-HRTF-based binaural headphone reproduction versus 4-CTC reproduction (music), where the positive slider range indicates a rating in preference of the 4-CTC system.

## Conclusion

In this paper, two variants of a headphone-based binaural reproduction were evaluated against a loudspeaker-based binaural reproduction in a dynamic listening experiment. A Wilcoxon rank sum test revealed no significant differences when comparing the median differences of a comparison involving the setup-HRTF-based binaural headphone reproduction versus the loudspeaker-based binaural reproduction to the median differences of a comparison involving setup-BRIR-based binaural headphone reproduction versus the loudspeaker-based binaural reproduction. Verification, however, was found when testing if the comparison of the setup-BRIR-based binaural headphone reproduction versus the loudspeaker-based binaural reproduction has a zero median difference using a two-sided sign test. This has evidence for the SAQI parameters *naturalness* (speech/music) and for *degree-of-liking* (speech/music) meaning that the setup-BRIR-based binaural headphone reproduction does not significantly differ from a loudspeaker-based binaural reproduction for these parameters.

## References

[1] Pulkki, V.: Spatial sound generation and perception by amplitude panning techniques. PhD Thesis, Helsinki University of Technology Laboratory of Acoustics and Audio Signal Processing, Helsinki, 2009

[2] Zotter, F.: Analysis and synthesis of sound-radiation with spherical arrays. PhD Thesis, University of Music and Performing Arts Graz, Graz, 2009

[3] Spors, S.: Active Listening Room Compensation for Spatial Audio Reproduction Systems. PhD Thesis, Technische Fakultät der Friedrich-Alexander-Universität Erlangen-Nürnberg, Nuremberg, 2005

[4] Masiero, B.: Individualized Binaural Technology. PhD Thesis, RWTH Aachen University, Aachen, 2012

[5] Blauert, J.: Spatial hearing: the psychophysics of human sound localization. The MIT Press, Cambridge, Massachusetts, 1997

[6] Huang, Y., Benesty, J., Clien, J.: Generalized crosstalk cancellation and equalization using multiple loudspeakers for 3D sound reproduction at the ears of multiple listeners. In ICASSP (2008), IEEE, 405-408

[7] Behler, G. K., Pollow, M.: Objective evaluation of the sweet spot size in spatial sound reproduction using elevated loudspeakers. The Journal of the Acoustical Society of America 123 (2008), 3614-3614

[8] Masiero, B., Fels, J.: Perceptually Robust Headphone Equalization for Binaural Reproduction. In Convention of the Audio Engineering Society 130 (2011)

[9] Parodi, Y. L., Rubak, P.: Objective evaluation of the sweet spot size in spatial sound reproduction using elevated loudspeakers. The Journal of the Acoustical Society of America 128 (2010), 1045-1055

[10] Parodi, Y. L., Rubak, P.: A Subjective Evaluation of the Minimum Channel Separation for Reproducing Binaural Signals over Loudspeakers. The Journal of the Audio Engineering Society 59 (2011), 487-497

[11] Lindau, A., Erbes, V., Lepa, S., et al.: A Spatial Audio Quality Inventory (SAQI). Acta Acustica united with Acustica 100 (2014), 984-994