Binaural resynthesis for comparative studies of acoustical environments

Alexander Lindau1, Torben Hohn1, and Stefan Weinzierl1

1 Dep. of Audio Communication, Technische Universität Berlin
Sekr. EN-08, Einsteinufer 17c, 10587 Berlin, Germany
tonalex@gmx.de, torbenh@gmx.de, stefan.weinzierl@tu-berlin.de

ABSTRACT

A framework for comparative studies of binaurally resynthesized acoustical environments is presented. It consists of a software-controlled, automated head and torso simulator with multiple degrees of freedom, an integrated measurement device for the acquisition of binaural impulse responses in high spatial resolution, a head-tracked realtime convolution software capable to render multiple acoustic scenes at a time, and a user interface to conduct listening tests according to different test designs. Methods to optimize the measurement process are discussed, as well as different approaches to datareduction. Results of a perceptive evaluation of the system are shown, where acoustical reality and binaural resynthesis of an acoustic scene were confronted in direct A/B comparison. The framework permits, for the first time, to study the perception of a listener instantaneously relocated to different binaurally rendered acoustical scenes.

1. INTRODUCTION

Auditory perception is often studied on the basis of comparative tests where listeners are asked to assess the difference between different stimuli. Such comparative setups are required in almost every domain of audio communication, from system evaluation, industrial sound design, planning of room acoustics and sound reinforcement systems up to basic research in psychoacoustics or empirical studies of media reception.

Whenever test listeners are supposed to be sensitive to small perceptual differences, an instantaneous switching the new stimulus is required. That is why field studies, e.g. when comparing different room acoustical environments, are largely inappropriate for methodical reasons. Hence, only the synthesis of acoustical environments by means of 3D audio technologies can allow for a direct confrontation of otherwise incompatible acoustical scenes. Since loudspeaker-based approaches such as Wave Field Synthesis (WFS) or Higher Order Ambisonics (HOA) require large-scale, hardly mobile installations, are affected by specific physical artefacts, and are still unable to (re)synthesize 3D soundfields with all their spatial properties, an approach based on binaural technology seems more appropriate, all the more when synthesis is not aimed at
a larger audience but for single listeners required in test situations.

Figure 1: Comparison of acoustical environments via binaural resynthesis

In the following, a comprehensive framework for the acquisition, post-processing and rendering of binaural representations of different acoustical environments is introduced, including a user interface for conducting listening tests. With the practical realisation of a framework as suggested in figure 1, the comparative study of otherwise incompatible situations will become accessible for the first time.

2. METHODS AND TOOLS

All binaural approaches towards (re)synthesizing a soundfield are based on the assumption that nearly all perceptual information used for orientation in our auditory environment is coded in the sound pressure at our eardrums. Hence, natural or artificial soundfields can be simulated on the basis of binaural room impulse responses (BRIRs), representing the acoustical transmission path from a sound source to the listener [1]. When convolving the measured BRIRs with an anechoic source signal the ear signals desired can be (re)created. When all system parameters such as the spatial resolution of the BRIR set, the spectral compensation of all involved transducers [2][3], and the latency of the dynamic auralization accounting for movements of the listener’s head [4][5][6] are carefully controlled, an auditory event can be created that is – even in direct comparison – largely indistinguishable from natural soundfields [7].

In the following we present a tool for comparative studies of acoustical environments based on binaural simulations, including

• a new head and torso simulator (HATS) for software-controlled, automatic measurement of complete sets of binaural impulse responses (BRIRs) in multiple degrees of freedom
• a standardized postprocessing of BRIRs
• a fast convolution algorithm allowing instantaneous switching between different BRIR data for multiple source locations
• a user interface for software-controlled listening tests based on different test design options

2.1. Acquisition

Several head and torso simulators for the acquisition of binaural impulse responses are available, such as the Kemar-maniquin [8], the Head Acoustics HMS-II/III series, the Bruel & Kjaer 4100 and the former Cortex (now Metravib) Mk1 as well as systems in the academic area such as the Valdemar-HATS [9] or the Aachen-head [10]. They are static systems with only limited possibilities to emulate movements of head, shoulders, and torso except by manual step-by-step reorientation [4]. A HATS-system developed at the TU Berlin in 2002 was the first to allow for an automated measurement of BRIRs for a (typical) horizontal rotation range of ±75°[11][12]. A similar system was presented recently for binaural assessment of car sound, consisting of a modified B&K 4100 head [13].

Figure 2: Design stages of FABIAN, close up of prototpe with automatable neck joint
The encouraging results of a perceptive evaluation of the TU System [8] gave rise to the subsequent development towards a more flexible and universal measurement system for the Fast and Automatic Binaural Impulse response Acquisition (FABIAN, Figure 2)[14][15]. It now allows for the precise, software-controlled horizontal and vertical orientation of an artificial head while measuring BRIRs using a multichannel impulse response measurement system with swept sine technique and noncyclic IR-deconvolution [16]. The whole device is mounted on a rotatable turntable that allows for the orientations of torso and head to be adjusted independently. Audio quality relating to parameters such as signal level, duration (FFT block size) and spectral coloration of the stimulus as well as the number of averages can be chosen to adapt the measurement to ambient conditions such as noise level or reverberation time. If input clipping is detected or if the signal-to-noise ratio (SNR) temporarily falls below a given threshold, a repetition of the last measurement is initiated. The simultaneous measurement of complete sets of BRIRs for up to 8 sources in high spatial resolution is controlled by a custom Matlab® application and thus considerably accelerated compared to conventional systems.

2.2. Postprocessing

Postprocessing so far includes the compensation of the headphone’s and microphone’s transfer function, an amplitude normalization, and a shortening of RIR lengths.

2.2.1. Headphone compensation

Although a multiplication of the BRIR spectra with the inverted complex transfer function (direct FFT deconvolution) of the mic-headphone-path seems an intuitive approach, two problems arise. First, the system’s transfer function is usually not minimum phase, therefore a stable and causal compensation filter may not exist [17]. Second, the highly nonlinear frequency response of headphones shows several narrow peaks up to +12 dB and dips with almost total extinction. Since their center frequencies may shift due to slight changes in the positioning of the headphone, a “perfect” equalization can lead to strong peaks with audible ringing artifacts.

This problem persists when the minimum and maximum phase part of the impulse response are inverted separately (homomorphic technique) [18], although an additional modeling delay applied to the maximum phase part provides an approximation of its acausal inverse, with filters as perfect as those derived from FFT-deconvolution and much smaller filter lengths.

Hence, another strategy was chosen. It is computationally simpler and superior in terms of remaining error energy.[18] The inverse filter is designed in the time domain with a least mean square (LMS) error criterion.

It can be shown [19] that the optimal filter, minimizing the LMS error of an non-perfect compensation and the “effort”, i.e. the energy of the compensation filter, is given by

\[ h = (C^T C + \beta B^T B)^{-1} C^T \delta (1) \]

Here, C is the filter to be compensated, \( \beta \) controls the absolute amount of regularization (i.e., non-perfect compensation) used and B is an FIR filter whose frequency response introduces a frequency weighting into the inversion process. Expression (4) uses matrix notation where C and B are convolution matrices of the respective filters, h and \( \delta \) are signal vectors, \( \beta \) is scalar. Regularization will be most effective in the passband of the filter which can be chosen without constraints on phase as only its energy will be effective [20]. To account for non-minimum phase components in the mic-headphone transfer function C a modelling delay is used. Therefore, the response of h is not designed towards reaching a ideal dirac pulse, but a delayed one, namely \( \delta(n-k) \) or \( \delta_k \). So the final expression used to design the inverse headphone filter shown in Figure 3 was

\[ h = (C^T C + \beta B^T B)^{-1} C^T \delta_k (2) \]

The lower cut-off frequency of the compensation is defined by the FIR filter length, in our case 1024 samples. As the compensation result depends not very strongly on absolute delay position, as long as it is not too close to the filters boundaries [18], the modelling delay was simply set to \( N/2 = 512 \) samples.

Since the optimal regularization strongly depends on the impulse response to be compensated, the parameters \( \beta \) and B can only be found by comparative listening [21].
In Figure 3 we have set $\beta = 0.8$ and $b(n)$ a highpass filter with -10dB in the stop band and a long transition range of 2-8 kHz. If a regularization on the LMS-error criterion is applied, the compensation of deep notches is affected first (Figure 3), whereas third- or sixth-octave band smoothing often used in FFT-deconvolution is less effective in this respect. The quality of the compensation reached was convincing even with a relatively short filter length of 1024.

![Figure 3: Frequency response of STAX SR 202, frequency response of inverse LMS filter (N=1024) with frequency-dependent regularization, calculated frequency response after compensation (from upper to lower)](image)

### 2.2.2. Length of BRIRs

A single set of BRIRs with 1°/5° horizontal/vertical resolution in a ±75°/±45° range and impulse responses of 2 s length sampled at 44.1 kHz in 32 bit floating point format creates a database of 2869 BRIRs allocating ca. 2 GB of data. Thus, data reduction strategies are crucial, as soon as several sets of BRIRs have to be held in random access memory of the convolution algorithm.

Here, several studies showed that a dynamic, head-tracked auralization is only necessary for the initial part of the BRIR, while the diffuse reverberation tail can be taken from a single direction without perceptual damage. For a small room with short reverberation time ($V = 185$ m³, $T_{30,1kHz} = 0.72$ s) Meesawat and Hammershi [22] found an offset of ca. 50 ms for the beginning of a diffuse reverberation tail. Since this offset is expected to increase with the mean free path of the room a similar listening test was repeated for BRIRs measured in a larger hall ($V = 10.000$ m³, $T_{30,1kHz} = 2$ s). A 3AFC test was done, using a short anechoic speech and drumset sample as test material. The BRIR sets were measured at a central seat position (third row) for a source placed on stage ca. 8 m away, which is roughly the critical distance of the room including the directivity of the source used. Listeners were asked to detect differences between the original 0°/0° BRIR and a concatenated BRIR using the diffuse part measured with the head turned left by 75°. An energy-preserving cosine window of 20ms was used for the crossfade. Whenever listeners correctly identified the concatenated IR, the offset was shifted in steps of 10 ms. After two runs with anechoic speech and a drumset sample of ca. 5 s duration (Run A, Run B) a third run with the diffuse tail from a source located 20 m away at 130°/30° (worstcase situation in [22]) followed (Run C). Here, only the the drum sample was used.

![Figure 4: Time offset inducing a just noticeable when concatenating the early part of a 0° BRIR with the diffuse tail of a 75° BRIR (A,B) and a BRIR with different direction and distance from the listener (C). Boxplots show median, interquartile range, and outliers](image)

Results for 23 subjects are shown in Figure 4. Median values for all subjects were 36 and 33 ms for sources with equal distance to the listener (run A/B) and 50 ms for sources with different distance to the listener. These offsets are close to the values determined in [22]. At the same time, there were subjects reliably detecting differences at offsets as late as 140 ms. Obviously, the effect of training (in our test introduced through a slow approach towards the JND) and “finding the right cue” plays an important role, as was reported by listeners reaching significantly different thresholds for different source signals.
Since the system is supposed to work for sensitive and trained listeners also, we use a block size of 16384 samples (~370ms) as dynamically refreshed direct part of the auralization. This not only reduces the size of the BRIR database, but is also used to increase the effective SNR of the simulation, because the BRIR used for the reverberation tail can be measured with a higher SNR resulting from several averaged measurements.

2.3. Realtime-Rendering

The realtime rendering application, running on Linux OS, is implemented as a JACK Audio server client[23]. It is able to render multiple sources at a time depending on computation power and RAM size. It uses fast nonuniform partitioned block convolution and a double static/dynamic caching algorithm to account for latency and memory limitations.

2.3.1. Implementation Details

The application is implemented in C++ and optimized for the x86 platform. Optimizing for PPC is quite easy though. Head movements of the listener as reported from the head tracking device are translated into OSC commands [24] by a separate application. For easy OSC integration the liblo open source library [25] was used.

The application manages a cache of impulse responses (Figure 5). While the maximum number of impulse responses stored in the cache can be configured, a separate cache manager thread watches the current headtracker position and loads impulse responses around the current head position (dynamic cache). As soon as the maximum number of responses is reached, it frees memory associated to the impulse response with maximum distance to the current position. A basic grid of impulse responses (static cache), loaded during program start can not be unloaded. The cache management allows transitions between different acoustical scenes even if the complete amount of BRIR data is too large to be held in random access memory.

To eliminate click artifacts, the exchange of impulse responses due to head movements is crossfaded in the time domain after convolution. The architecture allows to do the mixdown of the different sources as part of the complex multiplication in frequency domain, saving memory as well as memory bandwidth. The software currently uses the fftw3 open source library [26] for the FFT, while the complex multiplication is implemented in a vectorized form with SSE instructions.

2.3.2. Complexity

The algorithm can be divided into the input stages (mainly FFT), the complex multiplications, and the output stages (IFFT, crossfades, mix). The computation complexity for each of the stages depends on the number of source channels L, the partition size N_part, the length of the impulse response N_BRIR and the sampling rate f_s. Computation complexity for each stage of the algorithm is given by

\[
\text{output} = \mathcal{O}(f_s \cdot \log N_{\text{part}}) \quad (3)
\]

\[
\text{input} = \mathcal{O}(f_s \cdot L \cdot \log N_{\text{part}}) \quad (4)
\]

\[
\text{cplxmul} = \mathcal{O}(f_s \cdot L \cdot \frac{N_{\text{BRIR}}}{N_{\text{part}}}) \quad (5)
\]
For common values of $N_{\text{BRIR}} = 10^4 \ldots 10^5$ for reverberation times of 0.5...2.5 s and $N_{\text{part}} = 10^2 \ldots 10^3$ (see below) the computation costs are dominated by the complex multiplication, increasing proportional to the sampling rate and the number of source channels while decreasing inversely proportional to the partition size. On an Intel CoreDuo 2GHz Thinkpad with 2GB of RAM we could render up to 6 sources in $1^\circ/5^\circ$ (hor./ver.) resolution with impulse responses of length $2^{17}$ (3 s) and a block size of 256 samples.

2.3.3. Latency

To minimize the latency towards head movements, the minimum partition size of the convolution equals the audio processing block size, as set from inside JACK. The convolution implemented is overlap-add, so the latency of the source signal is one audio block. This latency is already imposed by the JACK audio server system. Head tracking is realized via a Polhemus Fastrack with 120 Hz update rate when using one single sensor. According to the following tabular overview with worst case latencies at 44.1 kHz sampling rate a reduction of block size below 128 samples would be largely ineffective considering the constant latencies of head tracker and tracker data interface.

<table>
<thead>
<tr>
<th>Blocksize:</th>
<th>128</th>
<th>256</th>
<th>512</th>
</tr>
</thead>
<tbody>
<tr>
<td>JACK Latency</td>
<td>2.9 ms</td>
<td>5.8 ms</td>
<td>11.6 ms</td>
</tr>
<tr>
<td>Serial Port</td>
<td>4.0 ms</td>
<td>4.0 ms</td>
<td>4.0 ms</td>
</tr>
<tr>
<td>Tracker</td>
<td>8.3 ms</td>
<td>8.3 ms</td>
<td>8.3 ms</td>
</tr>
<tr>
<td>Sum:</td>
<td><strong>15.2 ms</strong></td>
<td><strong>18.1 ms</strong></td>
<td><strong>23.9 ms</strong></td>
</tr>
</tbody>
</table>

Table 1: Audio output latencies for different block sizes

All sources are rendered continuously. For instantaneous switching between multiple sets of BRIRs OSC commands are sent from the graphical user interface of the listening test software switching audio outputs on and off (figure 6).

3. EVALUATION

To evaluate the plausibility of the binaural synthesis which is a pre-requisite for the validity of comparative studies as outlined in figure 1, a direct AB comparison of a natural and a resynthesized acoustic environment was done.

3.1. Listening Test Design

As a challenging setup, both in computational and in acoustical respect, a concert-hall-like environment was chosen with the auditorium maximum of the TU Berlin ($V = 10,000$ m$^3$, $T_{30,1kHz} = 2$ s). For two source positions and the measurement dummy seated in the third row, ca. 8 m from the sources, a measurement was conducted with $1^\circ/5^\circ$ horizontal/vertical resolution ranging from $\pm 75^\circ/\pm 30^\circ$, resulting in a database of 3926 BRIRs and lasting about 23 hours. Using a bass-emphasized, constant envelope log-sweep of FFT order 18 with 3 averages, the resulting BRIRs were saved with a length of 3.5 s (Figure 7). The overall SNR reached was about 95dB, with a frequency dependence shown in Figure 8.

The listening test was designed as a simple forced choice AB comparison test, where each subject listened to 80 randomized pairs of stimuli. Each pair consisted of a natural presentation (loudspeaker) and a binaurally simulated presentation of the same signal.
3.2. Results

The overall detection rate including all 2800 decisions (80 pairs à 35 subjects) was 52.9%. This is a small but tested on the basis of a Chi²-distribution with 2800 samples – still a statistically significant difference from the guessing rate of 50%. If we look at the detection rate per subject, we find a rather symmetrical distribution ranging from 32% to 72% for the 35 subjects tested (figure 9). Based on a Chi²-distribution for 80 samples the null hypothesis (“subjects are not able to identify the simulation”) has to be rejected for 8 out of 35 subjects (22.8%), those with detection rates of more than 59%.

After a training period each pair of stimuli was presented once, and the listener was asked to identify which stimulus was the simulation. The duration of each stimulus did not exceed 6s. Acoustically specifically transparent headphones (Stax SR 202) remained attached during the whole test. Accordingly, BRIRs measured with the headphones attached were used to account for the remaining shadowing effect of the external soundfield. Some conditions were varied as independent variables for statistical analysis. These included a. the degrees of freedom in head movement exploited by the rendering engine (horizontal vs. horizontal and vertical), b. different anechoic stimuli (male speech, female speech, acoustic guitar, trumpet, drums, taken from the EBU SQAM CD and the Archimedes collection) and c. the source position (0°/0° vs. 130°/30°).

35 subjects took part in this listening test, 12 female and 23 male. Attending a lecture about virtual acoustics and familiar with potential flaws of the technology they could be considered as “expert listeners”.

It is interesting that 5 of those 8 subjects in their questionnaire explicitly indicated that they were mostly guessing during the test. This suggests that they had mainly been better in remembering certain auditory cues and assigning them (rather) consistently to either reality or simulation, while the correct assignment was done merely by chance. This is also supported by the symmetry of the histogram in figure 9.

To find out which attributes were used by the subjects to discriminate between reality and simulation, a questionnaire was filled out right after the test. On the one hand, the answers were analysed as indications for potential flaws of the system. One the other hand the answers should reveal which features draw most attention when reality and simulation are directly confronted. Looking at the answers of those subjects who could (rather) consistently discriminate between the

![Figure 7: ETC of measured BRIR for left ear with head in 0° direction (left: frontal source, right: rear source). Note different direct sound levels but identical SNR.](image)

![Figure 8: Binaural room transfer functions and noise spectra, ipsilateral ear (left) and contralateral ear (right) for a frontal source (8m distance), head turned 75° left.](image)

![Figure 9: Frequency distribution of detection rates in AB-discrimination test, 5% margin for one-sided test is shown.](image)
two conditions (outside of the two-sided 5% Chi² values in figure 9), seven discrepancies were named at least twice. When ordered for frequency of occurrence, these were

1. spectral differences (5x)
2. difference in source localization (4x)
3. difference in reverberant energy (2x)
4. difference in energy on contralateral ear (2x)
5. difference in loudness (2x)
6. latency (2x)
7. change in tone during head movements (2x)

It is interesting that, although already considerable effort has been made to compensate the ear canal and headphone transfer function, spectral differences were still most obvious to the subjects.

Since localization differences were reported only with the 130°/30° source and not with the frontal source, they are most likely due to binaural features such as a slight mismatch in interaural time delays (ITD) for certain listeners due to the non-individual HRTFs used.

Concerning the general performance of the binaural simulation it should be kept in mind that the reported cues did not allow any listener to detect the simulation with a probability of more than 72%.

With regard to future listening tests the detection rate was also analysed per audio content, i.e. depending on the anechoic source signal used.

![Figure 10: detection rate depending on type of stimulus used](image)

As can be seen in Figure 10 the acoustic guitar sample (bourée by J.S. Bach) and the two speech samples seemed to be most suited to uncover potential artefacts of the simulation, while subjects were less sensitive when the drum sample and the trumpet sample was used. The guitar sound with a combination of transient and tonal features was obviously best suited to make slight differences in timbre as well as in localization audible.

The detection rate was slightly higher for the 0°/0° source than for the 130°/30° source (53.9% vs. 51.8%) and also when horizontal and vertical tracking was used (53% vs. 52.7%). While it is known that listeners are most sensitive to directional shifts close to the 0°/0° direction [27], the relevance of BRIRs in two degrees of freedom will be subject to further studies.

4. USER INTERFACE

For comparative studies of different acoustical environments the binaural rendering engine can be controlled by a specifically designed user interface realized in Matlab® and allowing for different listening test designs. The interface sends OSC commands to the rendering engine via TCP/IP (figure 6).

Test designs already implemented include

- AB comparisons of different binaurally synthesized or combined real vs. binaural reproductions such as the test presented above. They generate a randomized sequence of stimulus pairs and produce an output report for statistical analysis.

- a qualitative test design according to the repertory grid method (RGT) conducting an automated elicitation of individual attributes by randomized triads of binaural stimuli. Creating tables of personal constructs as a basis for subsequent quantitative ratings, the method has been successfully applied to study the perception of different multichannel recording and reproduction techniques [28].

- rating scales based on individually composed attributes, recommendations such as ITU-R BS 1284 [29], IEC 60268-13 [30], AES 20 [31] or semantic differentials resulting from a qualitative pretest as mentioned above.
5. APPLICATIONS AND OUTLOOK

The binaural framework presented will be an efficient tool whenever

- complete sets of binaural room impulse responses (BRIRs) have to be acquired in multiple degrees of freedom, in high directional resolution, and with high speed

and

- perceptive evaluations shall be based on instantaneous switching between different acoustical environments.

We see potential applications in the evaluation of the acoustical impression in different rooms and different listening positions, different configurations of sound reinforcement systems, or in a very fundamental confrontation of a natural acoustic environment, such as a concert hall, and its electroacoustic representation by different recording and reproduction systems. If the focus is extended to the binaural synthesis of computer-modelled environments other scenarios become accessible such as the design of sound installations of even the evaluation of historical developments such as the evolution of concert hall acoustics and its impact on perceptive attitudes [32].

Technical applications could include automotive audio assessment and the enhancement of speech intelligibility in teleconferencing and VoIP applications through binaural synthesis.

Whenever complete sets of BRIRs are acquired for multiple listening positions or even for a given listening area where listeners shall be allowed to move freely over a narrow grid of binaurally sampled listening positions, speed becomes an important issue. This applies to the speed of measurement already provided by the acquisition tool presented, but also to the speed of access to binaural data within the rendering application. Here, efficient strategies of data reduction have to be implemented. It is therefore essential to examine to what extent methods such as interpolation or principal component analysis, successfully applied for “lossless” compression of HRTF data [27][33], are equally efficient for binaural room impulse responses.

6. REFERENCES


[33] http://jackaudio.org