Sensory Profiling of Individual and Non-individual Dynamic Binaural Synthesis Using the Spatial Audio Quality Inventory

Alexander Lindau¹, Fabian Brinkmann¹, Stefan Weinzierl¹

¹Audio Communication Group, Technical University of Berlin, Berlin, Germany.

Summary
Data-based dynamic binaural synthesis (DDBS) aims at resynthesizing arbitrary acoustical environments by convolving measured binaural room impulse responses (BRIRs) with anechoic audio material and by playing back the result via headphones. Dynamic interaction is provided by observing the head movements of the listener and exchanging BRIRs accordingly and in real time. The perceptual accuracy of DDBS strongly depends on the used BRIRs being either individual or non-individual, i.e. reflecting the listener’s own or some alien morphology, e.g. of an artificial head and torso simulator (HATS). While the general perceptual accuracy has been assessed in terms of authenticity for individual and plausibility for non-individual DDBS with promising results, the future application of DDBS as an in simu reference, i.e., as a substitute of the acoustic reality in comparative evaluations, requires a qualitatively differentiated knowledge of its perceptual performance. Therefore, we asked nine subjects to rate the perceived differences when comparing exemplary state-of-the-art implementations of individual and non-individual DDBS directly to the acoustic reality. For sensory profiling we applied a questionnaire based on the ‘Spatial Audio Quality Inventory’ (SAQI), a recently proposed descriptive vocabulary for the perceptual evaluation of virtual acoustic environments (VAEs). Results revealed systematic deviations from acoustic reality for both the individual and the non-individual simulation, with the non-individual simulation exhibiting qualitatively more and quantitatively larger ones. Results are interpreted with respect to future applications of state-of-the-art binaural simulations.

PACS no. 43.66.+y, 43.55.+p

1. Introduction
Data-based dynamic binaural synthesis (DDBS), based on either individual or non-individual binaural room impulse responses (BRIRs), allows to resynthesize arbitrary acoustic environments with a high degree of realism. Hence, it has often been employed as an acoustic reference in listening tests. For example, Pellegrini [1, sect. 6.5] used non-individual DDBS to benchmark the perceptual transparency of a hybrid data/model-driven approach for the simulation of an acoustic control room. However, results obtained from such in simu evaluations might be misleading in cases where the reference simulation deviates systematically from the acoustic reality. Such deviations where observed for instance by Møller et al. [2]: While the authors found localization performance for (static) individual binaural recordings to be close to real life, an increased number of median plane confusion and distance errors was found for non-individual binaural recordings. Since localization performance alone might not be sufficient in proving the suitability of DDBS as a transparent acoustic reference, Brinkmann et al. [3] assessed the auditive authenticity (i.e., the indistinguishability from a given acoustic reference, here: the reality) of a state-of-the-art implementation of individual DDBS. Even this strictest imaginable criterion for acoustical simulations was could be met for some subjects and for a non-critical audio stimulus (male speech). For a state-of-the-art implementation of non-individual DDBS, the less strict criterion of plausibility (i.e. indistinguishability from expectations towards a corresponding real event) can be met [4]. Since assessments of the overall perceptual accuracy do not provide further information on remaining perceptual deficiencies, the current study reports preliminary results from an empirical assessment of qualitative and quantitative deviations perceived when directly comparing
individual and non-individual DDBS to the acoustic reality. In order to achieve both a sensory complete and practically relevant assessment of auditive differences the questionnaire used was based on a recently proposed descriptive consensus vocabulary for the evaluation of virtual acoustic environments, the Spatial Audio Quality Inventory (SAQI, [5]). Observed perceptual deficiencies are of fundamental interest both in terms of the performance of each approach individually, and in comparison. Furthermore, results allow drawing first conclusions about applying DDBS as a general acoustic reference simulation.

2. Methods

2.1. Listening Test Setup

Listening tests were conducted in the recording room of the Federal Institute for Music Research (SIMPK), Berlin ($V = 122 \text{ m}^3$, $RT_{1\text{kHz},\text{oct.}} = 0.65$ s). Subjects were seated on a chair with an adjustable neck rest and a small table for placing a computer mouse and a haptic MIDI interface, needed for conducting the binaural room impulse response (BRIR) measurements and the listening test. An active near-field monitor (Genelec 8030a) was placed in front of the subjects at a distance of 3 m and at a height of 1.5 m. With a critical distance of 0.8 m and a loudspeaker directivity index of ca. 5 dB at 1 kHz this setting should have resulted in a slightly emphasized diffuse field component at the listening distance. As an optical interface an LCD screen was placed at eye level and 2 m in front of the subjects, not obstructing the direct sound path from the loudspeaker (see Figure 1). For sound reproduction, a low-noise DSP-driven amplifier and extraaural headphones providing full audio bandwidth were used (BKSytem, [6]). Headphones were worn throughout the BRIR measurements and the subsequent listening tests, allowing for a minimally disturbed, instantaneous switching between binaural simulation and the corresponding real sound field (see [3], sect. 2.2 for an extended discussion). Head positions were controlled using a Polhemus Patriot head tracker.

2.2. Measurement of Non-individual BRIRs

Non-individual BRIRs and headphone transfer functions (HpTFs) were measured using the FABIAN head and torso simulator [7] with its built-in microphones located at the blocked ear canal at the bottom of the cavum conchae. During measurements FABIAN was seated on the chair similarly as a real subject while the BK211 headphones were fitted onto his head.

Figure 1. Listening test setup. The loudspeaker to the right of the subject was not used in the test reported here.

Then, BRIRs were measured for head-above-torsoro-orientations in a physiologically comfortable range of $\pm 34^\circ$ azimuth and with a resolution of 2$^\circ$ required for smooth rendering during head movements [8].

2.3. Measurement of Individual BRIRs

The measurement method for individual BRIRs was reported already in [3]; see there for a more detailed description of procedure and apparatus. In order to minimize effects of temporal drift, individual BRIRs and HpTFs were measured directly before each listening test. We used Knowles FG-23329 miniature electret condenser microphones flush cast into conical silicone earmolds [9] for measuring BRIRs at the blocked ear canal. Before starting the measurements, subjects put on the headphones and were familiarized with the procedure. First, subjects were asked to approach and hold a certain horizontal head orientation with the help of optical and acoustical guidance signals. Then, the measurement level was adjusted to be comfortable for the subjects while avoiding limiting of the DSP-driven loudspeakers or headphones. Finally, subjects themselves started each BRIR measurement by pressing a button on the MIDI-interface after they had moved their head to the target position and reached the latter within $\pm 0.1^\circ$. During a BRIR measurement, head movements of more than $\pm 0.5^\circ$ or $\pm 1$ cm would lead to a repetition of the measurement, which rarely happened. In this way, BRIRs were measured for the same angular range and resolution as described in sect. 2.2. Subsequently, ten individual HpTFs were measured per subject while rotating the head back and forth once between each measurement. After the measurements, which took about 30 minutes, the experimenter removed the microphones without having to remove the headphones. Sine sweeps of an FFT order 18 provided BRIRs with an average peak-to-tail SNR of approx. 80 dB.
without the need for averaging. All audio processing was conducted at a sampling rate of 44.1 kHz.

2.4. Headphone Equalization

Headphone compensation filters for both FABIAN and all individuals were designed using a weighted regularized least mean squares approach [10]. Individual filters were calculated based on the average of ten HpTFs measured per subject. PEQ-regularization of distinct notches in the HpTF as described in [9] was used to limit filter gains. The compensated headphones approached a target band-pass consisting of a 4th order Butterworth high-pass with a cut-off frequency of 59 Hz and a 2nd order Butterworth low-pass with a cut-off frequency of 16.4 kHz. For the individual DDDBS individual headphone filters were applied, whereas for the non-individual case a filter obtained from the FABIAN device was used [9].

2.5. Post Processing of BRIRs

Pre-delays were removed from BRIRs by means of onset detection. From these delays the interaural time differences (ITDs) were calculated and stored separately. During the listening test, ITDs were reestablished by using a real time variable delay line (see sect. 2.7), thereby efficiently reducing typical deficiencies of dynamic binaural rendering as, e.g., localization instability, latency, comb-filter and switching artifacts [11]. Then, BRIRs were normalized with respect to their mean magnitude response between 200 Hz and 400 Hz, and truncated to 44100 samples using a squared cosine fade out.

2.6. Loudness Matching

For the individual auralization, loudness matching was achieved by calculating a correction factor from the ratio of the RMS-levels of two recordings made at the subject’s ears when passing pink noise (1) through the loudspeaker, and (2) through the binaural simulation while the subject’s head was frontally aligned. For the non-individual auralization pinna cues differed between the ‘recording individual’ and the listener. Hence, a perfect loudness matching is difficult to attain. As a best case approximation, non-individual BRIRs were adjusted to exhibit the same RMS-level as the individual BRIRs.

2.7. Auralization

Dynamic auralization was realized using the fast convolution engine fWonder [7]. Real-time reinsertion of the ITD [11] was used for both individual and non-individual binaural rendering. Additionally, in case of non-individual BRIRs the ITDs were individually corrected based on the subjects’ head diameters [11]. fWonder was also used for applying the HpTF compensation filter and the loudspeaker target band-pass. The playback level was set to 60 dB(A).

2.8. Audio Stimulus

A pulsed pink noise (0.75 s noise, 1 s silence, 20 ms ramps) was used as stimulus, considered to be most appropriate to reveal potential flaws in the simulation. The bandwidth of the noise stimulus was restricted with a 100 Hz high-pass in order to restrict possibly audible variations due to low frequency background noise in BRIRs.

2.9. Spatial Audio Quality Inventory (SAQI)

The Spatial Audio Quality Inventory [5] was used to construct a questionnaire and rating scales for perceived deviations from reality in a qualitatively differentiated way. The SAQI is a consensus vocabulary comprising 48 verbal descriptors for auditive qualities considered to be relevant for the assessment of virtual acoustic environments. The vocabulary was generated by a Focus Group of 21 German experts for virtual acoustics. Five additional experts helped to verify the unambiguity of all descriptors and the related explanations. Moreover, an English translation was generated and verified by eight bilingual experts. The SAQI descriptors may be sorted into eight overall categories (Table 1) and are to be considered as ‘perceived differences with respect to [descriptor name]’.

<table>
<thead>
<tr>
<th>Category</th>
<th>Quality name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone color</td>
<td>Tone color bright-dark, High-/Mid-/Low-frequency tone color, Sharpness, Roughness, Comb filter coloration, Metallic tone color</td>
</tr>
<tr>
<td>Tonalness</td>
<td>Tonalness, Pitch, Doppler effect</td>
</tr>
<tr>
<td>Geometry</td>
<td>Horizontal/Vertical direction, Front-back position, Distance, Depth, Width, Height, Externalization, Localizability, Spatial disintegration</td>
</tr>
<tr>
<td>Room</td>
<td>Reverberation level, Duration of reverberation, Envelopment (by reverberation)</td>
</tr>
<tr>
<td>Time</td>
<td>Pre-/Post-echoes, Temporal disintegration, Crispness, Speed, Sequence of events, Responsiveness</td>
</tr>
<tr>
<td>Dynamics</td>
<td>Loudness, Dynamic range, Dynamic compression effects</td>
</tr>
<tr>
<td>Artifacts</td>
<td>Pitched/Impulsive/Noise-like artifact, Alien source, Ghost source, Distortion, Tactile vibration</td>
</tr>
<tr>
<td>General</td>
<td>(Overall) Difference, Clarity, Speech intelligibility, Naturalness, Presence, Degree-of-Liking, Other</td>
</tr>
</tbody>
</table>

Table 1. Spatial Audio Quality Inventory, English version. See http://dx.doi.org/10.14279/depositonce-1 for add. details.
SAQI attributes reflect ‘bottom-up’ as well as ‘top down’ perspectives of auditory perception related to specific aspects of VAE technology. Each descriptor is complemented by a short written clarifying circumscription and suitable dichotomous, uni- or bipolar scale end labels.

2.10. Listening Test

The German version of the SAQI was used for the listening test. Three items ("Speed", "Sequence of events", "Speech intelligibility") were omitted as they were thought to be of minor relevance for the current test. The remaining 45 items were administered to the subjects using the free Matlab® listening test software whisPER2. The presentation order of individual/non-individual simulations was randomized across subjects in a balanced fashion, with one condition to be assessed completely before switching to the next. The presentation order was randomized within each SAQI category (cf. Table 1), as well as for the SAQI categories themselves. During the listening test, each SAQI item was assessed individually, starting with a presentation of the written circumscription to remind subjects of the exact meaning of each descriptor. Then, a rating scale was presented together with two play-buttons for immediate comparison of (hidden) binaural simulation and reality. If no difference was perceived with respect to a specific quality, a rating could be skipped.

2.11. Listener Panel and SAQI Training

Nine subjects with an average age of 30 years (6 male, 3 female) participated in the listening test. No hearing anomalies were reported. Subjects received written circumscriptions for all SAQI qualities before visiting the lab. Ambiguous items were discussed with the experimenter on site. Subjects were instructed to actively exploit head movements when assessing auditive differences and to compare stimuli as long as they wanted.

3. Results

Nine subjects rated both individual and non-individual simulations with respect to 45 auditive qualities in a fully repeated design summing up to $2 \times 9 \times 45 = 810$ individual ratings. Skipping a quality was treated as a zero rating. Ratings were pre-screened using boxplots with outliers, individual profile plots, and qualitative statements. For the individual simulation, one subject skipped the complete questionnaire indicating that no difference was perceived. Another subject reported a ringing artifact after accidentally touching the headphones. All ratings were included in the subsequent statistical analysis.

Ratings were tested for normality using the Shapiro-Wilk test. Violations of normality at $p \leq 0.2$ were observed for almost 80% of the SAQI items and results are thus reported in a non-parametric way using boxplots (Figure 1). In order to highlight a potentially systematic effect within one condition, boxes were shaded when the inter-quartile range (IQR) did not include zero. Potential differences between tested conditions are indicated by non-overlapping neighboring IQRs. A Wilcoxon signed-rank test proved the ratings of the (overall) ‘difference’ item to be significantly different between the two conditions ($p = 0.012$). At three occasions subjects used the item “Other” to explain differences they felt not to be covered by the SAQI. These included an impression of ‘phasiness’, a ‘resonating, sustaining’ impression, and a ‘more room-like, more diffuse, more enveloped’ simulation, each mentioned by a different subject.

4. Discussion

Figure 1 shows the results for both conditions and averaged across subjects. Three qualities were never perceived: ‘front-back confusion’, ‘pre-echoes’ and ‘alien source’. For non-individual DDBS one item (‘impulsive artifact’), for individual DDBS six items (‘roughness’, ‘metallic tone color’, ‘envelopment’, ‘ghost source’, ‘distortion’, ‘tactile vibration’) were never mentioned. While initially, these ‘non-ratings’ may interpreted positively as the simulation being indeed perceptually transparent with respect to these qualities, it should be carefully considered whether the used stimulus was truly appropriate to elicit difference ratings. However, as nearly all items were addressed at least under one condition, this concern appears to be of minor relevance.

About half of the items, where differences were reported, showed some inter-individual variation but only minor systematic offsets from reality, with IQRs including zero. For the non-individual simulation, these aspects may be due to inter-individual morphological variability, whereas for the individual simulation they reflect limits of measurement accuracy (both physiologically and physically).

As indicated by shaded IQRs in Figure 1, the non-individual DDBS was perceived as notably different from reality with respect to 19 qualities, whereas the individual DDBS differed in only 12

---

2 http://dx.doi.org/10.14279/depositonce-31
aspects. Qualities with larger deviations can be found in most of the SAQI categories including — for the non-individual simulation — tone color and tonalness (attenuated mid frequencies, comb filter, a perception of tonalness/increased pitch/pitched artifact), geometry (a reduced externalization and localizability, an increased depth, width and spatial disintegration), time and general (reduced crispness, clarity, naturalness, liking). For individual DDBS, strong deviations were reported only with respect to a reduced sharpness, loudness, clarity and (spatial) presence and an increased depth of the auditory event. These aspects have to be considered as most problematic when aiming at using DDBS as an acoustic reference simulation. However, and especially for non-individual simulations, they might particularly benefit from future improvements, too.Offsets in horizontal direction were not considered here as these were most probably due to a scaling bias: Results were not reproducible after changing the scaling method (direct reporting in degree instead of using a ±180° slider).

5. Conclusion

Results of our study revealed potentially systematic offsets of non-individual binaural simulations from reality, including attributes related to coloration, reduced localizability and distance errors, which may be attributed to deviating pinnae cues. For individual binaural simulations, only few minor deviations with respect to spectral coloration and geometry were observed. These deviations have to be considered when binaural synthesis is used to provide an acoustic reference simulation: Especially, when assessing binaurally simulated sound fields ‘as is’ in an absolute fashion, i.e. without referring to an explicitly given reference simulation, any deviation induced by the simulator itself will bias the assessment. In contrasting, if certain binaurally simulated sound fields are to be judged in comparison to similarly binaurally simulated reference sound fields, deviations of the simulator might be tolerable, as the effect can assumed to be constant under all tested conditions.

Considering the differentiated picture of perceptual properties, the Spatial Audio Quality Inventory (SAQI) has proven to be an informative and comprehensive measuring instrument for the differential diagnosis of dynamic binaural synthesis.

This investigation was supported by a grant from the German Research Foundation (DFG WE 4057/3-1).

References


Acknowledgements
Figure 1. Comparative plot of deviations of non-individual/individual data-based dynamic binaural synthesis from individual acoustic reality as rated on the Spatial Audio Quality Inventory. Ratings are displayed as interquartile boxes (IQR), medians, and remaining data points. IQR-boxes are shaded when not including zero. Ratings for “front back confusions” are given in percentages as in this case the used scale was dichotomous (“Yes/No”). Square brackets around abscissa labels indicate qualities not rated by any subject. Ratings are coded to enable a most intuitive interpretation with positive/negative values encoding and increased/reduced perception of the respective auditory quality in the binaural simulation under test. Ratings of ‘hor./ver. direction’ are most probably exaggerated due to scaling bias (see article text).