INFLUENCE OF SPATIAL COMPLEXITY AND ROOM ACOUSTIC DISPARITY ON PERCEPTION OF QUALITY FEATURES USING A BINAURAL SYNTHESIS SYSTEM

Stephan Werner¹, Florian Klein¹, and Karlheinz Brandenburg¹

¹ Technische Universität Ilmenau, Ilmenau, Germany

ABSTRACT

This contribution presents investigations on the influence of scene complexity and room acoustic disparity on the perception of different quality features using a binaural headphone system. The quality features “spatial presence” and “listener envelopment” are investigated next to “perceived externalization” and “localization” of an auditory event. The test uses three different rooms with distinct room acoustic characteristics and several scenes with different spatial complexity. The work addresses the question if the quality features can profit by the different audio scenes or not. The results show that spatial presence is influenced by spatial complexity while room acoustic disparity influences listener envelopment. Furthermore, externalization and localization are not affected by spatial complexity regardless of the personalization method used for binaural synthesis.

Index Terms— binaural synthesis, spatial presence, listener envelopment, externalization, localization

1. INTRODUCTION AND MOTIVATION

The presented work is motivated by the quality formation process proposed by Jekosch [1] and Raake [2]. This process describes the creation of quality of experience of multimedia systems and services as a process of comparison and judgment between desired and perceived quality features. Figure 1 shows an extension of this quality formation process. The extension contains not the formation process by itself. It is a model for the assumption that the multimedia system with its context of use influences the buildup of quality of experience. Furthermore, feedback to the system enables an adaptation of the quality of system to reach plausible auditory illusions for example.

This work is part of a research project to measure the influence of different context dependent quality parameters on the perception of an auditory scene using binaural synthesis. The investigated parameters are audio-visual divergence [3], personalization of the system [4], acoustic divergences [5], spatial complexity of the presented scenes and room acoustic disparity. The named parameters are assumed as relevant for a binaural synthesis system to create a plausible auditory illusion in the context of use.

2. BINAURAL SYNTHESIS SYSTEM

For the test stimuli, binaural recordings of individual and artificial binaural room impulse responses (BRIRs, using a KEMAR head and torso simulator) for the selected rooms, sound sources and positions have been done. The binaural system is personalizable to reduce within-cone and out-of-cone of confusion errors [6] and to increase the fidelity of the simulations compared to real loudspeakers [7]. Different rooms have been chosen to include different room acoustic properties including reverberation and source-receiver distances. Reverberation encourages the perception of externalization of an auditory illusion and the impression of distance [8]. The headphones are equalized using individual headphone transfer functions (HPTFs) if individual BRIRs are used. HPTFs from the head-and-torso simulator are used if artificial BRIRs are used. In-ear microphones are used to measure individual BRIRs and HPTFs at the entrance of the blocked ear canal of each test person. The microphones are not removed between the BRIR and HPTF measurements. The inverse of a HPTF is calculated by a least-square method with minimum phase inversion [9]. Stax Lambda
Pro headphones which fulfill the requirements for open headphones are used for playback [10].

3. QUALITY FEATURES UNDER TEST
In the experiments we asked for four quality features which we assume to be relevant in binaural synthesis of a spatial audio scene ([11], [12], [13], [14], [5] and others). For this work we define the features as follows. Localization describes the perception of direction of an auditory event in the horizontal plane. A correct localization of auditory events is assumed to be a desirable feature when comparing virtual loudspeaker setups with real setups. This is true especially for binaural simulations of multi-channel loudspeaker setups as the ones proposed in MPEG-H [15]. Similar localization accuracies are investigated in this work. Envelopment describes the perception of the position of an auditory event outside or inside the head of the listener [17], [18], [5]. We define the perception of an event very close to the head or ears as in-head-localized or non-externalized. Envelopment is a crucial feature to reach a plausible spatial auditory illusion with binaural headphone systems. Results from several investigations show that perceived externalization is significantly lower if listening and synthesized room are divergent [17], [19], [5]. The Envelopment of Reverberation [14] or listener envelopment [11] describes the perception of reverberant sound surrounding the listener. A high energy of late reflections and a long reverberation time increase perceived sound surrounding the listener. A high energy of late reflections and a long reverberation time increase perceived sound. A high energy of late reflections and a long reverberation time increase perceived sound surrounding the listener. A high energy of late reflections and a long reverberation time increase perceived sound surrounding the listener. A high energy of late reflections and a long reverberation time increase perceived sound surrounding the listener. A high energy of late reflections and a long reverberation time increase perceived sound surrounding the listener. A high energy of late reflections and a long reverberation time increase perceived sound surrounding the listener.

4. OBJECTS OF INVESTIGATION
4.1. Spatial complexity
The test conditions include several audio scenes with different arrangements of one target sound source and one, two or three noise sound sources. Figure 2 gives an overview about the audio scenes. Scene A consists of one sound source as target sound source (TS) at different directions. Scene B has one lateral, frontal, or rear target sound source (TS) and one variable noise sound source (NS1). Scene C has one target sound source (TS) and two lateral noise sources (NS1 and NS2).

Audio scenes in front and back of the listener are used to investigate quadrant errors in the localizations task. The positions of the sound sources are determined by a direction angle for azimuth and a radius. An angle of 0° is chosen for elevation. The possible horizontal directions are located at 60° steps on a clock-like circle with a radius of 2.5 m from the listening position. A detailed description of the angles of the target and noise sources is given in table 1 for the different scenes.

Table 1. Audio scenes and source directions used in the listening test; see also figure 1; TS=target source, NS=noise source.

<table>
<thead>
<tr>
<th>Scene</th>
<th>Front angle</th>
<th>NS1 angle</th>
<th>NS2 angle</th>
<th>Rear angle</th>
<th>NS1 angle</th>
<th>NS2 angle</th>
</tr>
</thead>
<tbody>
<tr>
<td>A0</td>
<td>0°</td>
<td>-</td>
<td>-</td>
<td>A1</td>
<td>180°</td>
<td>-</td>
</tr>
<tr>
<td>B0</td>
<td>0°</td>
<td>180°</td>
<td>-</td>
<td>B1</td>
<td>180°</td>
<td>0°</td>
</tr>
<tr>
<td>C0</td>
<td>0°</td>
<td>180°</td>
<td>-</td>
<td>C1</td>
<td>180°</td>
<td>0°</td>
</tr>
<tr>
<td>A30</td>
<td>30°</td>
<td>-</td>
<td>-</td>
<td>A30</td>
<td>150°</td>
<td>-</td>
</tr>
<tr>
<td>B30</td>
<td>30°</td>
<td>270°</td>
<td>-</td>
<td>B30</td>
<td>150°</td>
<td>270°</td>
</tr>
<tr>
<td>C30</td>
<td>30°</td>
<td>330°</td>
<td>-</td>
<td>C30</td>
<td>150°</td>
<td>330°</td>
</tr>
</tbody>
</table>

Audio scenes are created by mirroring at the inter-aural axis to cover the cone-of-confusion errors (see table 1). Audio scenes in front and back of the listener are used to investigate quadrant errors in the localizations task. The positions of the sound sources are determined by a direction angle for azimuth and a radius. An angle of 0° is chosen for elevation. The possible horizontal directions are located at 60° steps on a clock-like circle with a radius of 2.5 m from the listening position. A detailed description of the angles of the target and noise sources is given in table 1 for the different scenes.

The scenes are assumed to be sparse scenes in the sense that especially localization and separation of the single sound sources is possible by humans. We are interested to find whether different source directions and number of sources within the synthesized rooms have a supporting or limiting effect on the perception of quality features.
4.2. Room acoustic disparity
Three different rooms with distinct room acoustic characteristics are used in the test. The rooms are a listening lab (HL; compliant to Rec. ITU-R BS.1116-1, V=179 m³), an empty seminar room (Hu201, V=224 m³), and an empty lecture hall (Leonardo DaVinci, LdV, V=766 m³) at TU Ilmenau. The rooms are different in several acoustic parameters like reverberation time, direct-to-reverberation energy ratio, inter-aural correlation, and others. As a summarized distinction the time behavior of reverberation is gives as energy decay curves in figure 3. The LdV has a longer reverberation while the HL has the shortest one.

![Fig. 3. Energy Decay Curves of the used rooms measured at 0° position with omnidirectional microphone.](image)

Loudspeakers (Genelec 1030A) are placed at each hour position of a clock-like circle with 2.5 m radius in each room to measure BRIRs. Individual BRIRs of the test persons are measured in the listening lab while artificial BRIRs using a head and torso simulator (KEMAR) are measured in all three rooms.

5. LISTENING TEST
Figure 4 gives an overview about the combinations of listening room and synthesized rooms for the listening test. Only the listening lab (HL) is used as listening room in the test. Therefore the combinations are “HL in HL”, “LdV in HL”, “Hu201 in HL”. The test stimuli consist of combinations of the three recording rooms, three audio scenes with 36 sub scenes and individual BRIRs and artificial BRIRs.

![Fig. 4. Combinations of listening room and synthesized room within the listening tests; HL=listening lab, LdV=Leonardo DaVinci lecture hall, Hu201=seminar room.](image)

The listening test is divided into sub-tests. The first sub-test is a listening test to rate the quality features localization and externalization within a single-stimulus test design. The binaural synthesis of different audio scenes and three rooms using individual measured and dummy head BRIRs is presented. Rating of localization of the auditory event is realized by indicating the direction on a rating sheet. The rating sheet is a top-down view with a symbolized human head in the middle. Several regions are selectable to rate perceived direction and externalization similar to [5]. The perceived localization and externalization is conducted for the target source only. Rating of externalization of the auditory event is realized by indicating inner, middle, or outer regions on the rating sheet. The following definitions are used in the test: a) midpoint: “The auditory event is entirely in my head and very diffuse.” b) inner circle: “The auditory event is entirely in my head and easy to locate.”; c) middle circle: “The auditory event is external but it is next to my ears or head.”; d) outer circle: “The auditory event is external and easy to locate.”; e) outer cloud: “The auditory event is external and very diffuse”.

The second and third sub-test is a listening test to rate spatial presence and listener envelopment within a reference free multi-stimulus test design. Listener envelopment is rated for different room acoustic disparities while spatial presence is rated for different spatial complexities. Binaural synthesis using artificial BRIRs is used for both tests. The whole auditory scene should be rated in these tests in contrast to the first test.

A female and two male speech signals with 6 s duration are used as audio signals. The female speech is the target source for test A. The male speech signals are the noise sources. Dutch speakers are chosen to avoid familiarization with the content of what is said.

6. RATINGS
Twenty-three test persons (three female and 20 male) with a mean age of 27 years (SD=3 years) participate in the listening test A and B. 70% of the persons report experience with perceptual tests. Fifteen test persons (two female and 13 male) with a mean age of 28 years (SD=3 years) participate in test C. All test persons report experience with perceptual tests.

6.1. Externalization
The ratings of the quality feature externalization are counted as frequencies. An externalization index is calculated as ratio between the number of ratings for “extern” (outer circle and outer cloud at the rating sheet) and all number of ratings. An index of “0” indicates an entire in-head-localization, while an index of “1” indicates full externalization of the auditory event. Figure 4 shows the externalization indices for scene A, B, and C of the listening lab (HL) for individual and artificial BRIRs.

The ratings show an increase of the externalization index for personalized binaural synthesis compared to KEMAR use. Lower indices are especially visible for 0° and 180°
direction. The externalization indices are comparable with indices from former experiments [5]. The different investigated scenes show no significant influence on perceived externalization (at $\alpha=5\%$, Wilcoxon signed-rank test). Furthermore, no significant differences are found regarding the number of sound sources and their spatial arrangement.

Fig. 4. Externalization indices with 95% confidence intervals for scenes A, B, C for different source directions and using KEMAR and individually recorded BRIRs of the listening lab (HL).

6.2. Quadrant errors
The ratings for the quality feature localization are counted as angles. The angle resolution is given with 30° by the hour positions on the rating sheet. For further analysis a quadrant error index is calculated as ratio between the number of confusions on the inter-aural axis and the total number of ratings. An index of “0” indicates no quadrant errors, while an index of “1” indicates total confusion of directions. Figure 5 shows the quadrant error indices for scene A, B, and C of the listening lab.

Fig. 5. Quadrant error indices with 95% confidence interval for scene A, B, C for different source directions and using KEMAR and individually recorded BRIRs in the listening lab (HL).

The ratings show an increase of the number of quadrant errors for 0° and 180° compared with other directions for non-personalized binaural synthesis. Less quadrant errors show up for personalized binaural synthesis compared with KEMAR use. These results are in line with other investigations on localization [6], [7], [5]. Like for the quality feature externalization, no significant influence of the number and spatial arrangement of audio objects on the occurrence of quadrant errors are found.

6.3. Spatial presence
A reference free multi-stimulus test design is applied to compare four different audio scenes with each other. Each comparison consists of the three audio scenes (A, B, C) for the different source directions. An additional low spatial quality signal (scene A in free-field conditions) is presented as anchor. The ratings are done on a presence scale from “0” (less present) to “100” (more present). No significant differences are found for the different source directions ($\alpha=5\%$, Wilcoxon signed-rank test). For further analysis the ratings for the different directions are combined. Figure 6 shows the ratings as median with 25 and 75 quantile for the different audio scenes, number, and spatial arrangement of sound sources.

Fig. 6. Spatial presence as median and 25/75 quantile for different audio scenes and number of sources; * sig. difference at $\alpha=5\%$ (Wilcoxon signed-rank test); HL=listening lab, FF=free-field.
Low interquartile distances (IQDs) are found for the low spatial quality anchor (“FF_A”). The test persons show a high inter-rater reliability for this signal type. The highest IQDs are found for the audio scene A (“HL_A”). No significant difference is found between the different scenes if the audio objects within a scene are at the same position (with label “_4”). In contrast, significant differences between the scenes are found for lateral arranged audio objects (“_2”, “_3”) and for the number of audio objects (e.g. “B_2”, “C_2”). Less spatial presence is rated for scene C with two noise sources at the same position and the target source at opposite position compared to scene B with only two sources at opposite position (“B_1”, “C_1”). Spatial presence is increased if the number of sound sources is increased and if the sources are spatially arranged around the listener.

6.4. Listener Envelopment
A similar reference free multi-stimulus test design as for the rating of spatial presence is used for the ratings of listener envelopment. The test compares one scene in all rooms. Additional audio scenes are created for free-field conditions as low spatial quality anchor. Listener envelopment is expected to be low because room information is missing in free-field conditions [20]. Ratings are on an envelopment scale from “0” (less enveloped) to “100” (more enveloped). No significant differences are found for the different source directions (at $\alpha=5\%$, Wilcoxon signed-rank test). For further analysis the ratings for the different directions are combined.

Figure 7 shows the ratings as median with 25 and 75 quantile for the different audio scenes, rooms, and number of sound sources. The ratings show that the quality feature listener envelopment changes with the presented room but not by the presented audio scene. In conformance with other studies for concert hall acoustics and virtual acoustics using wave field synthesis, listener envelopment is increased if the amount of energy from late reflections and the reverberation time is increased [20], [21]. An investigation from Nowak et al. [21] shows that a decrease of an ensemble width from $12^\circ$ to $4^\circ$ at a frontal position has no influence on perceived listener envelopment. This statement can be confirmed and extended. Listener envelopment is not effected by different source directions and scenes all-around the listener.

7. CONCLUSION AND OUTLOOK
The influence of context dependent quality parameters on the formation of quality of experience has been investigated. Because of the multiple relations between the quality features and context dependent parameters a proposed description space is given in [22]. The space is classified according to context dependent quality parameters, quality features, and adaptation as the coordinate axes. The intersections represent the results of the single perceptive experiments as derived statistics. As preliminary statement we can summarize that spatial complexity in our tests had no influence on externalization, localization, or listener envelopment but it does have an influence on spatial presence. More spatially distributed scenes increase spatial presence. The investigated room acoustic disparity shows an influence on the perception of listener envelopment. More reverberant rooms are rated with higher listener envelopment. The arrangement of the different audio scenes on frontal, rear, or lateral directions has no significant influence on the perception of listener envelopment and spatial presence.

The results of the experiments for localization and externalization are within the accuracy of the tests inline
with former tests despite of different test designs and conditions ([5], [6], [7], and others). No significant influence of the investigated scenes and number of audio objects on the occurrence of quadrant errors and on perceived externalization is found. However, the scenes used in this test are sparse. Further investigations on dense scenes in which localization and separation of the single sources is much more difficult seem to be necessary. It remains unclear how spatial scenes are perceived if there is no localizable single audio object.

In summary, the tests reported in this paper and earlier experiments show very good reliability of the quality features under test. These features seem to be adequate to describe spatial perception in the context of a binaural synthesis system. However, perceived spatial presence clearly can be viewed as a more dimensional quality feature. If the audio scene in the test is sparse (first scene), we find less consistent ratings across the test persons. The test persons report in an interview after the test that for this scene other features like localization, externalization, spaciousness, and envelopment maybe rated. Spatial presence seems to be an accumulation of other sub-features like the above-mentioned ones and maybe others.

8. ACKNOWLEDGMENT

We thank all test participants for the participation in the tests and interest in research. Furthermore, we thank Martina Böhme for her assistance and the course advanced psycho acoustics summer term 2014. This work is supported by a grant of the Deutsche Forschungsgemeinschaft (DFG Grant BR 1333/14-1).

9. REFERENCES