Listening experiment on the plausibility of acoustic modeling in virtual reality

Kajetan Enge[1], Matthias Frank[2], Robert Höldrich[3]

[1] *University of Music and Performing Arts Graz, Austria, Email: kajetan@enge.at*

[2] *Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Austria, Email: frank@iem.at*

[3] *Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Austria, Email: hoeldrich@iem.at*

Introduction

Without its sound, physical space itself would feel implausible to the people in it. Plausibility is described by Kuhn-Rahloff as the matching between a situation and an inner reference that is dependent on individual experiences [1]. Pellegrini describes the difference between plausibility and authenticity. He states that authenticity occurs when a real environment is reproduced in all its physical features, while plausibility occurs when the important aspects of perception are properly modeled [2]. In acoustic terms, some of these aspects can be externalization, localization, timbre, reverb, and dynamics of a source in a certain environment. In general, feeling "present" in a virtual reality situation seems to be easier when spatial acoustics are involved, especially when those are rendered dynamically [3, 4, 5, 6].

This study aims to test different virtual acoustic modeling techniques for their perceived plausibility in virtual reality (VR). To do so, a listening experiment in a virtual environment was done with 20 participants. They were asked to enter a virtual room with the help of an HTC VIVE head-mounted display (HMD[1]) and a pair of open headphones [7]. In the virtual room, their task was to rate the plausibility of the acoustic modeling of a virtual loudspeaker. This loudspeaker was placed at a central position in the room and the participants were able to move around it. Through the HMD, the participants saw a virtual version of the room they also were physically in. The dimensions of the virtual and the real room were aligned such that they were able to move freely in the whole space without anything getting in their way. In the real room, there also was a real loudspeaker at the same central position. Figure 1 shows the virtual room with the virtual loudspeaker. The questions we were investigating with this experiment are: (I) How are different strategies of acoustic modeling performing in a VR situation? (II) Is there an influence of the degrees of freedom on the perceived plausibility in a VR situation?

Setup

This listening experiment required to develop a test environment in VR. The requirement for its design was to provide a fully dynamic acoustic model that can be entered in virtual space so that the participants can move freely through the room with six degrees of freedom (DoF) while receiving the corresponding audio signals. The test environment should be flexible enough to offer the researcher a variety of acoustic models for investigation.

[1] `https://www.vive.com/eu/product/#vive%20series`

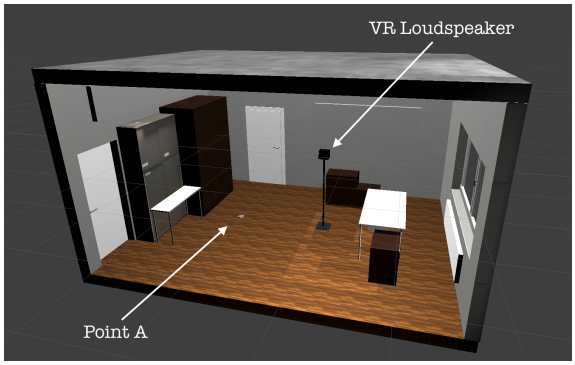


Figure 1: The virtual studio with the dimensions 6 m × 4.4 m × 3 m [x/y/z] and the loudspeaker at (0.4 m, −0.2 m) from the center. 'Point A' for BRIR measurement and as position for static and rotation-only listening at is (−1.4 m, −0.25 m). This room size also seems to be at the limit of the possible area covered by a VIVE tracking system with two base stations.

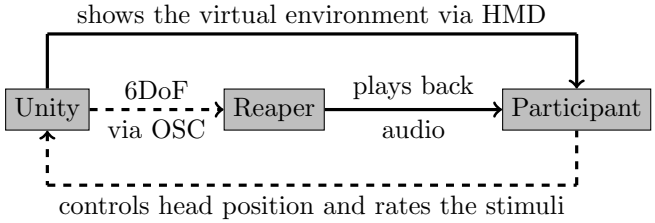


Figure 2: The structure of the experiment: Unity sends the 6 DoF controlled by the head position of the participants to Reaper where the corresponding audio is rendered and played back via the open headphones or the real loudspeaker.

For this purpose, the IEM Plug-in Suite[2] was used. It offers a set of Plug-ins to work with Ambisonic [8] signals up to seventh order. The audio rendering was done in the digital audio workstation Reaper[3]. The visuals were designed in Unity[4], a 3D modeling software, that allowed us to easily connect the HMD. Since Unity always knows the exact position and orientation of the participant's heads, it was possible to provide Reaper with the six DoF data via open sound control (OSC[5]). Reaper then used this information to binaurally decode the acoustic model to a set of open headphones. In the experiment, we studied the rated plausibility of six different acoustic models and the real loudspeaker, in combination with two visual conditions and three levels of DoF.

[2] `IEMPluginSuite:https://plugins.iem.at`
[3] `Reaper:https://www.reaper.fm`
[4] `Unity:https://unity.com`
[5] `OSC:http://opensoundcontrol.org`

Acoustic models
Besides the real loudspeaker, the six acoustic models were the following: 7th, 3rd, and 1st-order Ambisonics including the loudspeaker's directivity pattern, 7th-order without the loudspeaker's directivity pattern, 7th-order with less reverb, and 7th-order direct sound with a convolved BRIR reverb that is independent of the orientation and position of the participant. The audio signal used in the experiment was male, English speech from [9], the EBU's "Sound quality assessment material, recordings for subjective tests."[6]

The acoustic virtualization consisted of an image-source model (ISM) for the first 236 reflections and a Feedback-Delay-Network (FDN) for the diffuse reverb. As the sound quality of an FDN depends on the number of channels it uses in the network, the FDN always employed a 64×64 matrix and was encoded at 64 almost equally spaced points into 7th order and decoded with the respective order. A measurement of the directivity pattern of the real loudspeaker was done to implement a 3rd-order directivity pattern for the virtualization. Also, a binaural room impulse response of the real loudspeaker in the real room was measured at position 'A' in front of the loudspeaker with the KU100 artificial head by Neumann[7]. The real loudspeaker was a Behritone C50A with a single driver and axisymmetric layout. Both these features were helpful for the modeling of its directivity pattern. The loudspeaker was positioned slightly off-center at 1.6 m height to prevent perfectly symmetrical left and right early reflections.

In the case of the 7th-order model, the virtual loudspeaker emits its sound with the measured directivity pattern. The discrete reflections are rendered with the ISM and the diffuse reverberation is done with the FDN. The 64-channel Ambisonic signal is rotated according to the head movements of the participants. In the end, this signal is decoded binaurally to the pair of open headphones. The *3rd Order* and the *1st Order* stimuli are modeled in the same way but with reduced Ambisonic orders. In the binaural room impulse response model *BRIR reverb* only the direct sound is dynamically modeled with 7th order. All the reverberation is done with a convolution of the speech signal with the reverberant part of the static BRIR of the studio room. In the virtualization, the reverb signal was delayed by 440 samples to compensate for the distance to the loudspeaker in the measurement and the delay that the Room Encoder causes at this distance. The *No Directivity* condition is equivalent to the 7th-order model, but the loudspeaker emits its sound with an omnidirectional pattern. As this condition brings more energy into the virtual room, the level of the reverb increases. As compensation, all the discrete reflections from the ISM were attenuated by 4 dB. The *BRIR & No Directivity* condition uses the BRIR reverb and the omnidirectional pattern for the encoded direct sound.

[6] https://tech.ebu.ch/publications/sqamcd
[7] Neumann:https://de-de.neumann.com/ku-100

Visual conditions
There were two visual conditions within the experiment. In the first one "Blind", the HMD showed a black picture to the participants. This stimulus aims to make this study comparable to known literature about binaural reproduction like [10]. The second condition showed the virtual version of the studio room the participants physically were in.

Degrees of Freedom
Within the DoF, we distinguished between three levels: Static, Rotation and Translation. 'Static' meant that a participant had to stand at the position in front of the loudspeaker marked on the virtual floor with the letter 'A', see figure 1. This position was 1.8m in front of the loudspeaker, at the same position where the static BRIR was measured. The participants were asked to look only in the direction of the loudspeaker without moving their heads. The condition 'Rotation' meant that the people should also stand at position 'A', but this time they were allowed to move their heads. The third possibility 'Translation' allowed to move freely in the virtual room. Whenever the next stimulus was a static or rotational one, the visual interface told the participant to walk back to position 'A' before they continued. With the visual condition "Blind" only the DoFs 'Static' and 'Rotation' were tested.

Method

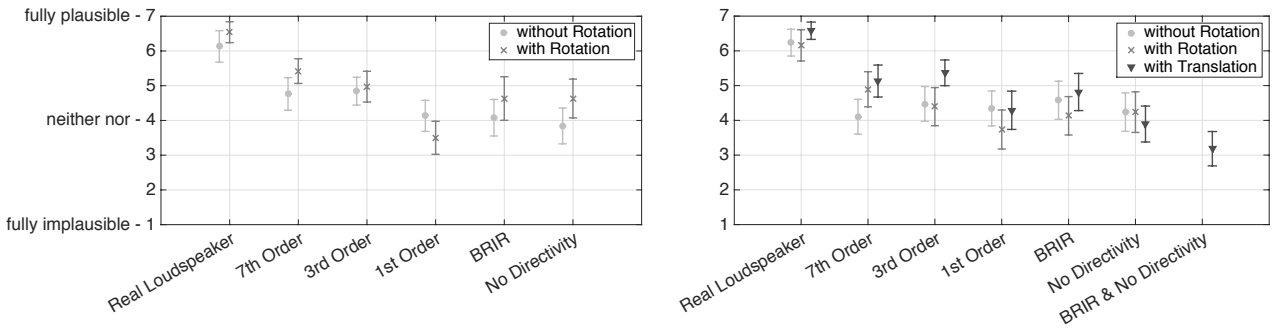
Instructions to the participants
To make sure everybody had the same concept of "plausibility" we provided a definition. The participants were instructed to ask themselves the following three questions every time they rate a loudspeaker:

1. "Is this audiovisual presentation plausible?"
2. "Is what I hear consistent with what I see?"
3. "Could a loudspeaker in such a room sound like this?"

To help the participants answer these questions, they were asked to pay attention to certain characteristics:

1. Externalization - "Does this sound like the loudspeaker is outside my head?"
2. Localization - "Can I hear the loudspeaker from the same direction I see it?"
3. Distance - "Do the acoustic and visual distances to the loudspeaker match? Does it sound closer or more distant than it looks?"
4. Timbre - "Does the tone color of the loudspeaker change depending on my position in the room accordingly to my expectations?" - Keep in mind that a real loudspeaker speaker does not sound the same from all directions.
5. Reverberation or room impression - "Do my acoustic and my visual room impression match? Does the room sound as big as it looks?"

The user interface
The participants had to rate the plausibility of the audio-visual presentation on an integer, Likert-type scale from 1 (= fully implausible) to 7 (= fully plausible).



(a) Without visuals ("blind"): Degrees of freedom are 0 and 3.

(b) With studio visuals: Degrees freedom are 0, 3, and 6.

Figure 3: 95% confidence intervals around the average ratings for different acoustic stimuli and degrees of freedom.

With a rotational movement of a VR-controller, a visual interface stepped through the possible ratings. To give their answers, the participants confirmed the current number by pressing a button on the controller. Also, the integer scale was defined for the positions 7, 4, and 1 before the experiment started:

- 7 - fully plausible: "It would be possible that a real loudspeaker in such a room would sound like this."
- 4 - neither-nor: "I realize that the virtualization doesn't fit very well, but I don't think it's fully implausible"
- 1 - fully implausible: "The sound and the picture do not match at all, or at least one of the above characteristics is completely different from what I expect."

When the participants rated one stimulus, the audio playback stopped and they were instructed to prepare for the next one by walking back to position 'A'. All the stimuli appeared in an individual, random order with the one constraint, that all the "blind" stimuli were rated first. This way, the participants didn't have to find their way back to position 'A' in the dark.

Results

20 mostly experienced participants took part in the study, one of them was excluded from the analysis due to bad reliability. All participants rated every stimulus twice, those are interpreted as independent ratings because the inter-rater standard deviations and the intra-rater standard deviations are similar.[8] The data is analyzed using two strategies. First, we study the influence of the different acoustic models on the perceived plausibility. Then we will have a look at the influence of the DoF. These analyses are done for both visual conditions. Just for the illustrations in the figures 3(a) and 3(b), but not for statistical analysis, we interpret the rank scale as an interval scale and therefore do not present median values, but average values with the corresponding CI. However, CIs of median values would also be misleading due to their strong discretization, therefore we decided to use average values on the plot.

[8]This paper investigates an excerpt of the conducted experiment, in which the participants rated also other combinations of visual and acoustic stimuli.

For the statistical analysis, Wilcoxon signed-rank tests [11] were evaluated. The cumulation of the alpha error was considered via a Holm-Bonferroni correction [12]. Cliff's Delta (referred to as Δ) [13] and Cohen's d (referred to as d) were used as measures for effect size. Cohen's d is technically not a valuable effect size measure for ordinal data, however, it is used here as an orientation for the reader because Cliff's Delta is less common. Furthermore, Cohen's d is only referred to when the Jarque-Bera-Test indicates that the involved data is normally distributed. Both Cliff's Delta and Cohen's d are considered as absolute values. Cohen's d is calculated according to [14] with pooled sample standard deviations. In the figures 3(a) and 3(b), the ordinates range from 1 to 7 and represent the ordinal scale of plausibility. The abscissas show the different acoustic conditions.

Blind - Ranking of acoustic models

Figure 3(a) shows that, with and without head-movement, the participants rated the real loudspeaker significantly more plausible than all the models ($p \leq 0.017$, $\Delta \geq 0.54$). We assume that most participants identified the real loudspeaker and then rated with a high score automatically. Without head-movement, the acoustic virtualizations are not significantly different from each other with only one exception between the *3rd Order* and *No Directivity* ($p = 0.019$, $\Delta = 0.36$, $d = 0.72$). When the participants were allowed to move their heads, they rated the *1st-Order* model significantly less plausible than *7th Order*, *3rd Order*, and *No Directivity* ($p \leq 0.027$, $\Delta \geq 0.37$, $d \geq 0.67$). This agrees with the findings in [10].

Blind - Differences in the degrees of freedom

The rotation tends to improve the results for all stimuli except for the *1st Order*. The differences are significant for the following stimuli: The real loudspeaker ($p = 0.008$, $\Delta = 0.14$), the *7th Order* ($p \leq 0.01$, $\Delta = 0.27$, $d = 0.53$), the *1st order* ($p = 0.045$, $\Delta = 0.236$), and *No Directivity* ($p = 0.013$, $\Delta = 0.2735$, $d = 0.49$). The significant deterioration in 1st order is due to the fact that the perceived distance to the source is noticeably reduced during a 90 degree rotation. It is interesting to note that rotation improves plausibility for 7th-order playback, while it degrades the 1st-order playback by emphasizing its weakness. For 3rd order, plausibility is not effected by rotation.

Studio - Ranking of acoustic models
Also here the real loudspeaker always performs significantly better than all of the virtualizations ($p \leq 0.005$, $\Delta \geq 0.51$).The *7th Order* and the *3rd Order* did not show any differences in plausibility, regardless of the degree of freedom ($p \geq 0.203$, $\Delta \leq 0.16$, $d \leq 0.31$). Hence, in this virtual reality environment, the plausibility of the 7th-order model was already achieved with 3rd-order modeling. Even without using dynamic reverberation in the acoustic model, the *BRIR* virtualization does not perform significantly differently from most others. The only significant difference here is to *BRIR & No Directivity* with $p \leq 0.001$ ($\Delta = 0.54$, $d = 1.06$). It seems to be important for plausibility that the direct sound is modeled dynamically. A dynamic reverb seems to be less important for a plausible impression. With translation, the *1st Order* and the *7th Order* are only weakly significantly different, however this comparison still has a medium effect ($p = 0.058$, $\Delta = 0.3$, $d = 0.55$). The possibility to move freely in the room improves also the plausibility of the *1st-Order* model. If the virtual loudspeaker employs an omnidirectional pattern, it sounds the same from all directions. In comparison to the other acoustic conditions this doesn't affect plausibility for 'No Rotation' and 'With Rotation' ($p \geq 0.181$, $\Delta \leq 0.21$, $d \leq 0.4$). Anyway, if the participants move freely in the room, the missing directivity pattern leads to a significantly worse plausibility when compared to the *7th Order* and the *3rd Order* ($p \leq 0.013$, $\Delta \geq 0.44$, $d \geq 0.84$).

Studio - Differences in the degrees of freedom
In the virtual studio, the degrees of freedom influence the *7th Order*, *3rd Order*, and *1st Order* just like without visuals. First, let us compare the DoFs 'No Rotation' and 'With Rotation'. In the case of *7th Order*, we see a significant difference with $p = 0.033$ and medium effect size ($\Delta = 0.28$). Within the *3rd Order*, the rotation does not change the resulting plausibility ($p = 0.856$, $\Delta = 0.003$, $d = 0.05$). Within the *1st Order*, even though figure 3(b) shows lower values for the rotation than without rotation, there is no significant difference ($p = 0.27$, $\Delta = 0.2$, $d = 0.037$). Translation improves plausibility for the *7th Order* in comparison to 'Without Rotation' ($p = 0.003$, $\Delta = 0.4$, $d = 0.7$). For the *3rd Order*, translation improves plausibility in comparison to both 'No Rotation' and 'With Rotation' ($p = 0.005$, $\Delta \geq 0.32$, $d \geq 0.67$).

Conclusion
In this study, we developed a test environment in virtual reality to investigate different modeling techniques for virtual acoustics. With current technology, we can model virtual acoustics in VR with 3rd-order Ambisonics without significant loss in plausibility compared to a 7th-order model. In practice, this results in a reduction from 64 to 16 Ambisonic channels, which can be helpful in both, streaming and for gaming engines. We showed that rotation and translation mostly increase plausibility and that source directivity is important in the direct sound. However, position-dependency of reflections seemed to be not important in our test scenario. Discussions with participants suggest that they were able to identify the real loudspeaker and then automatically answered with a very high rating. A possible side-effect of this is a downgrading for all of the acoustic virtualizations.

Acknowledgments
The authors thank all listeners for their participation in the experiment.

References
[1] C. Kuhn-Rahloff, "Prozesse der Plausibilitätsbeurteilung am Beispiel ausgewählter elektroakustischer Wiedergabesituationen," Ph.D. dissertation, Technische Universität Berlin, 2011.

[2] R. S. Pellegrini, "Quality assessment of auditory virtual environments." Georgia Institute of Technology, 2001.

[3] B. Jens, *Spatial Hearing: Psychophysics of Human Sound Localization.* MIT Press, 1997.

[4] C. Hendrix and W. Barfield, "The sense of presence within auditory virtual environments," *Presence: Teleoperators & Virtual Environments*, vol. 5, no. 3, pp. 290–301, 1996.

[5] M. Naef, O. Staadt, and M. Gross, "Spatialized audio rendering for immersive virtual environments," in *Proceedings of the ACM symposium on Virtual reality software and technology*, 2002, pp. 65–72.

[6] D. R. Begault and L. J. Trejo, "3-D sound for virtual reality and multimedia," 2000.

[7] N. Meyer-Kahlen, D. Rudrich, M. Brandner, S. Wirler, S. Windtner, and M. Frank, "DIY Modifications for Acoustically Transparent Headphones," in *AES 148th Convention, e-Brief 61*, 2020.

[8] F. Zotter and M. Frank, *Ambisonics.* Springer, 2019.

[9] EBU, "Sound quality assessment material, recordings for subjective tests," *EBU-SQAM*, Oktober 2008.

[10] M. Zaunschirm, M. Frank, and F. Zotter, "BRIR synthesis using first-order microphone arrays," in *Audio Engineering Society Convention 144.* Audio Engineering Society, 2018.

[11] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945. [Online]. Available: http://www.jstor.org/stable/3001968

[12] S. Holm, "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979. [Online]. Available: http://www.jstor.org/stable/4615733

[13] N. Cliff, "Dominance statistics: Ordinal analyses to answer ordinal questions," *Psychological bulletin*, vol. 114, no. 3, p. 494, 1993.

[14] J. Hartung, G. Knapp, and B. K. Sinha, *Statistical meta-analysis with applications.* John Wiley & Sons, 2011, vol. 738.