Audio Engineering Society

Conference Paper

Presented at the AES International Conference on
Audio for Virtual and Augmented Reality
2020 August 17 – 19, Virtual

Towards Transfer-Plausibility for Evaluating Mixed Reality Audio in Complex Scenes

Stefan A. Wirler[1], Nils Meyer-Kahlen[1], and Sebastian J. Schlecht[1,2]

[1]*Acoustics Lab, Dept. of Signal Processing and Acoustics, Aalto University, Espoo, Finland*
[2]*Media Lab, Dept. of Media, Aalto University, Espoo, Finland*

Correspondence should be addressed to Stefan A. Wirler (stefan.wirler@aalto.fi)

ABSTRACT

The evaluation of mixed reality audio is typically approached under the paradigms of either authenticity or plausibility. While the first refers to the identity of a real and a virtualized sound source, the latter measures the degree of belief in cases where no direct reference is available. We refer to transfer-plausibility as the ability of a virtualized source to stand alongside multiple real sound sources. We present a perceptual experiment where listeners detect and identify a sound source as being virtualized using dynamic non-individualized binaural rendering under varying scene complexity. Scene complexity is controlled by a varying number of loudspeakers. We demonstrate that the presented methodology mitigates ceiling effects, typically encountered in authenticity and plausibility tests.

1 Introduction

In mixed reality (MR), the aim is to supplement the real world with virtual objects by convincingly reproducing the perceptual stimuli via sensory displays. Apart from the problem of delivering a convincing visual representation, the sound of virtual objects should seamlessly blend into the real acoustic environment.

A physically accurate recreation of virtual sound sources is an ideal solution for MR; however, in many cases, practically unfeasible. Thus, when approaching this task, it is most challenging to clarify the perceptual requirements. The hardest goal is *authenticity* [1], which is commonly defined as the perceptual identity of a real sound and its virtualization. Experiments designed to test authenticity should comprise discriminative tasks. To test whether a difference in the representations can be perceived at all, headphones are worn during the presentation of the real and the virtual sound. This was done in [2], using an ABX paradigm, which is a 3 Interval - 2 Alternative Forced Choice (2AFC) test, and in [3], using a 3AFC test. Also, the oddball design in [4] allows for direct comparison. To achieve authenticity in this sense, all auditory cues must be identical. Practical rendering systems often have minor, but easily identifiable differences, which lead to ceiling effects. Also, in some studies, the aim is to compare different impairments.

As a less demanding alternative to authenticity, the concept of *plausibility* of virtual sound sources was introduced. Plausibility refers to the degree of similarity to an internal reference [5] or the "agreement with the

listener's expectation towards an equivalent real acoustic event" [2]. In practical applications, frequently, no direct comparison between real and virtual sound is available, which makes plausibility a more relevant measure of quality. One option is to assess plausibility using an ordinal scale [6, 7]. As an alternative, a model that relies on immediate real/virtual answers has been introduced, which utilizes signal detection theory to identify individual sensitivity and response bias [8, 9], which is a major advantage of such a design. In plausibility experiments, it is important to minimize the amount of comparability between the trials, e.g. by varying the spatial location of the sources, the signals, or their spectral content.

Both authenticity and plausibility have severe disadvantages as a measure of reproduction quality. Authenticity allows any perceptual difference to contribute to the identification rate. Most notably, minute spectral deviations can be detected by listeners (as low as 1 dB [10]) similar to those introduced by headphone re-positioning, HRTF individualization, reproduction latency and calibration methods. We argue that not all such deviations are central issues of spatial reproduction quality. On the other hand, plausibility ratings are highly dependent on the individual predispositions such as professional training, expectation, and task interpretation [5].

In many application scenarios, a third situation is much more common, in which similar, but not identical, real and virtual sound sources exist next to each other, thereby introducing a certain amount of controlled comparability. We refer to this with the notion of *transfer-plausibility*. In this work, we argue that transfer-plausibility can be operationalized as a clearly defined task which allows the rating of a wide range of reproduction qualities. In our experiment, up to eight real and virtual sound sources are presented simultaneously in a echoic room. The task is to identify the loudspeaker which was virtualized over headphones. Here, the *scene complexity*, i.e., the number of concurrent loudspeaker signals, is varied to adjust the task's difficulty. The following paper is structured as follows. In Section 2, we present the technical setup of our experiment. Section 3 presents the testing methodology of transfer-plausibility and scene complexity. Section 4 gives the results of our study.

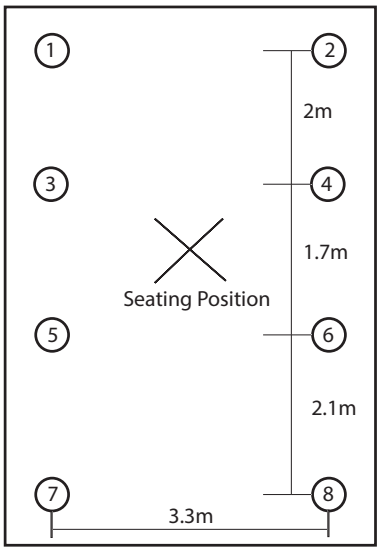


Fig. 1: Layout of the room used for the experiment. Circles indicate loudspeakers and the cross indicates the seating position.

2 Technical Setup

In this section, we describe the methods used for the virtualization of the sound sources. It should be noted that we did not intend to achieve a perfect binaural reproduction, as we are aiming to investigate practically feasible methods and the influence of the *scene complexity*. An accurate reproduction of virtual sound sources would depend on the use of individualized head-related transfer functions (HRTFs) [11, 12, 3] and position-dependent binaural room impulse responses (BRIRs), rendered dynamically via head-tracking [12, 13, 14]. Also, individualized compensation of the Headphone Transfer Function (HpTF) should be included. In our experiment, we only used a single static BRIR per loudspeaker to model the room reverberation. The direct path is however virtualized dynamically, using non-individualized HRTFs with the SPARTA binauralizer [15]. In the current version, the default setting uses numerically modeled HRTFs. The dynamic positional change is taken into account with a 6-degree-of-freedom (6DoF) head tracking system which offers high precision and low latency [16, 17].

2.1 Room Setup

The experiment was conducted in the Immersive Sound Studio (ISS) at Aalto University. The room has a size

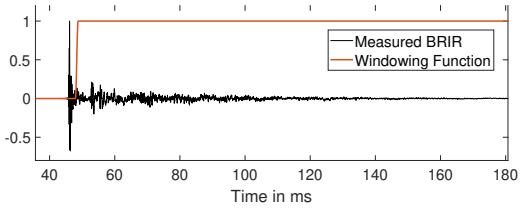
of 6m x 3.60m x 3.20m and is equipped with eight loudspeakers (Genelec 1030A) on ear level. Its reverberation time is about 0.4s and therefore it provides a more realistic acoustic environment (e.g. similar to a living room), when compared to the anechoic space. Figure 1 shows the placement of the loudspeakers as well as the subject's seating position during the listening test. In the beginning of every trial, the listener is facing the front wall behind loudspeaker 1 and 2.

2.2 BRIR Synthesis

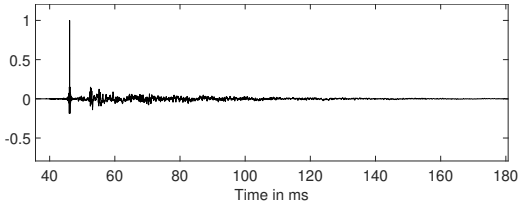
The measurement of the BRIRs was performed once before the experiment. The measurement was conducted with a human listener, who was seated in the listening position. The BRIRs were captured with binaural microphones[1] at the entrance of the ear canal. The headphones were worn during the measurement to incorporate their influence. Exponential sine sweeps with a duration of 3s, generated by the ITA-Toolbox [18], were used as the measurement signal and played and recorded via an RME Fireface UCX at a sampling frequency of 48kHz. A single BRIR was measured for each of the eight loudspeakers, while the listener was facing the loudspeaker. In the reproduction, only the early reflections and the reverberation tail of the measured BRIRs were used (see Fig 2a). Further, the BRIRs were not adjusted to directional and positional changes of the listener's head.

The direct sound is the signal portion most sensitive to adjustments of head orientation and position in real-time, while the reverberation is less position-dependent. Therefore, we remove the direct sound of the measured BRIR and replace it with a dynamically generated synthetic direct path. The direct path is removed from the measured BRIR by multiplying a step-function window with a rising sine-squared shaped fade-in, as shown in Figure 2a. The direct-to-reverberant ratio is crucial to preserve as it is indicative of the perceived source distance [19]. For this, we determine the energy of the removed direct sound of the measured BRIR and adjust the energy of the direct sound generated by the binauralizer accordingly. Additionally, the BRIR tail is shifted in time to align with the generated HRTFs. During real-time processing, the stimulus is then convolved with the HRIR (Binauralizer in SPARTA [15]) and with the wet BRIR in parallel (Real-Time Multiconvolver in HISS-Tools [20]). The head orientation and

[1] Sound Professionals - Low Noise In-Ear Binaural Microphone



(a) Windowing of the measured BRIR (left channel) to remove the direct sound.



(b) Processed BRIR (left channel) with combination of synthetic generated direct sound and addition of early reflections and tail.

Fig. 2: Example of a measured and a processed BRIR.

the position change of the head is taken into account by the tracking system and fed into the binauralizer. These two signal components are then added to generate the headphone output. An example of a synthesized response is shown in Figure 2b.

2.3 Headphones

The choice of headphones is a crucial element in conducting experiments for the comparison of real and virtual loudspeakers. In such an experiment, the influence of the headphone on the real sound reproduced by the loudspeakers should be kept as low as possible. In terms of passive solutions, the design should be as open as possible, while still maintaining an acceptably flat HpTF. In the present study, a modified version of the AKG K702 headphones was used. The modification is carried out by the remodeling of the ear cushion, details can be found in [21].

To compensate for the HpTF, headphone equalization filters were derived. Although repeated individual measurements of the headphone transfer function would be preferable [22], the additional effort for the participants was avoided by using a robust, non-individual equalization filter. The filter for the modified model was derived based on measurements with the KU100 dummy head, which does not feature an ear-canal simulation and thus corresponds to the employed BRIR

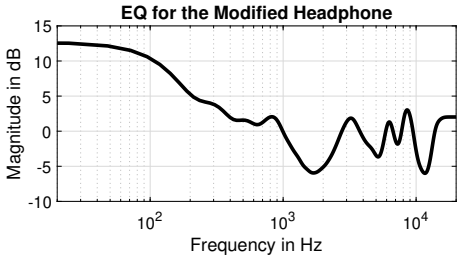


Fig. 3: Minimum Phase Headphone Compensation.

measurements, c.f. Section 2.2. The HpTF measurement was repeated twenty times for both headphone orientation (turn headphones). In between these measurements, the headphones were re-positioned, to cope with variability due to displacement. After half the measurements the headphone was flipped. After averaging, the obtained frequency responses were set to be constant below 50 Hz, to avoid high amplification, and above 15kHz, because the high variability due to replacement the headphone prevents a meaningful filter design. Afterwards, the responses were smoothed using a 1/6-octave filter and minimum-phase impulse responses were derived using cepstral windowing. The IRs were cut to 256 samples, using a squared sine window for fade-out. The overall signal is then convolved with the derived HpTF by the HISS-Tool Real-Time Multiconvolver (based on the convolution without input/output delay described in [23]).

3 Perceptual Test Methodology

3.1 Experiment

In the experiment, three different stimuli were used - music, speech, and pulsed pink noise. To assess the effect of the complexity of a scene, one source is virtualized. Four different test cases, with one, two, four or eight sources, were carried out using two different types of reproduction qualities. A virtual source was either modelled only using dry HRTFs or with the combination of HRTFs and BRIRs (Section 2.2). Additionaly, the situation of only real loudspeakers playing exists (none of the loudspeakers is virtual). The experiment was split into a training session, a session with HRTF rendering and a session with BRIR rendering. The BRIR session also included the case of only real sources playing. The training session consisted of 10 trials. The participant was presented with the different stimuli, the different qualities and the different test cases. In the training session, the participant was given the correct answer as well as which rendering quality was used. Before the start of each session, the participant was informed about the used rendering method. Both of the sessions consisted of two trials for every combination of the rendered and the real cases (test cases, stimuli, rendering and two trials $4 \times 3 \times 2 \times 2 = 48$ trials, resulting in an overall of 96 trials). The stimuli were looped and participants were allowed to listen to one trial as long as they liked. During the test, the subject was seated in the center of the room, but was allowed to move their head freely during the test. The signals played by the speaker as well as the position of the playing speaker was randomized for every trial. A demonstration video of the experiment is available online[2].

3.2 Test Cases

The different test cases represent the scene-complexity. They differ in the number of active loudspeakers. For every trial, a random subset of loudspeakers was selected and one of the signals was assigned to each of them. The aim of this procedure was to minimize the opportunity for inter-trial comparison. Within the experiment, the positions of the active speakers were indicated to the participants. The first test case has the lowest complexity, as only one loudspeaker, either virtual or real, is active. In the second test case, two loudspeakers are active simultaneously. This gives the participant three different answering possibilities. Either the first or the second loudspeaker is virtual or none of the loudspeaker is virtual (only real loudspeakers are playing). The scene complexity is then raised for the third and the fourth test case, in which four or eight of the loudspeakers are playing, respectively. This increase of scene complexity results in 5 or in 9 different answer possibilities, respectively.

3.3 Stimuli

For the music stimuli, a different instrument was reproduced by each loudspeaker. The music/instruments were taken from the song "What's Trumps" by "Rhythmussportgruppe" [24]. The instruments chosen were: Piano, Acoustic Guitar, Drums, Cow Bell, E-Base, Saxophone, Trumpet, and E-Guitar. The speech signals

[2] https://www.sebastianjiroschlecht.com/publication/SceneComplexity/

consisted of samples in four different languages (Danish, English, German and French) spoken by a male and female person, respectively. For the noise stimuli, pink noise was used. Its amplitude was shaped by a squared sine function, with different frequencies (1-3Hz). In contrast to other studies [2, 3], the stimuli are not high- and lowpass filtered. The position of the stimuli (music instrument, speaker, noise frequency), was randomized for every trial in the experiment. The signals are looped during the whole experiment. Attention was paid to the seamless looping of the signals.

3.4 Subjects

The experiment was conducted with 15 participants, whereas three were female and twelve were male. Every participant is experienced with dynamic binaural reproduction as well as in music. Therefore, the participants can be acknowledged as experienced listeners.

4 Results

For plausibility studies with only one source, Signal Detection Theory is a useful tool for comparing sensitivity d' and bias β of the individual participants [8]. In the case of more complex scenes in the present design, this task turns into a simultaneous detection and identification problem [25], governed by a similar model. Yet with the current amount of data available, we look at the results more in the sense of classification, using confusion matrices and the associated terminology. Note that this is a multiclass classification problem and that an answer is only counted as a true positive, if the presence of a virtual source was detected, and the source was correctly identified. An answer is considered true negative, if the absence of a virtual source was detected correctly. The accuracy is given by the proportion of correct answers in total (see Fig. 4).

4.1 Effect of Stimulus

In authenticity studies, the type of stimulus had a substantial effect on the result, e.g., [2]. It was found that pulsed pink noise significantly aids the listener in discriminating between real and virtual sound sources. In such direct comparisons, minor spectral deviations become most audible for broadband stimuli. In the present study, we were also able to observe this effect.Fig. 4a and 4c presents the mean percentage of correct answers, averaged over all participants, for HRTF and BRIR rendering, respectively. Only in the case of BRIR rendering a trend is observable for virtual sources. Listeners can recognize more reliably if the stimulus is pulsed pink noise (68%) compared to Music (43%) or Speech (41%). This could be caused due to coloration effects caused by the BRIR rendering. In the case of HRTF rendering, the opposite effect is seen for the case of only real sources.

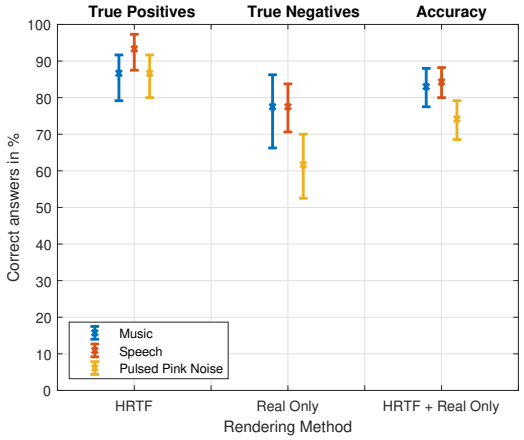
4.2 Effect of Scene Complexity

Fig. 4b and 4d present the rates of correct answers for the four complexity levels, i.e., for one, two, four and, eight loudspeakers. A central observation is the percentage of virtual sound sources that were correctly identified, i.e., true positive rate. Fig. 4d indicates that the true positive rate decreases with increased scene complexity. For a single active source, the task is equivalent to previous studies, as the identification is effectively a decision on the source being either real or virtual. When only HRTFs were used for rendering a single source, all participants were able to identify the virtual speaker. With increasing scene complexity, however, even the sub-optimal pure HRTF rendering leads to a reduced number of correct identifications. Therefore, presentations of more complex scenes can reduce ceiling effects even for low-quality rendering.

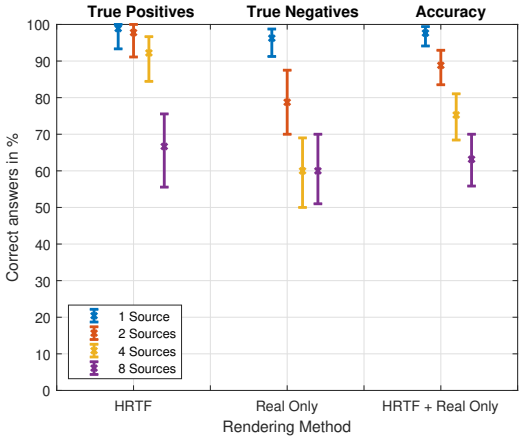
Incorporating the non-individualized BRIR leads to an overall reduction of true positives, see "True Positives" in Fig. 4d. The mean percentage of correctly identified virtual sources is between 40% and 61%, depending on the scene complexity. Note that the chance probability depends on the number of presented sources. For a single source, the chance level is at 50%. For scene complexity 4, either none or one of eight virtual sources were possible, it is reduced to $\sim$11%. Thus, for BRIR rendering, the true positive rate is at chance for the single source, but above it for the most complex scene.

Interestingly, the participants identified a virtual source more reliable in the case of two sources compared to only one source (see Fig. 4d). In the case of two sources, the participant had a direct comparison of a virtual and a real source. Therefore, it seems it is easier to distinguish with a direct reference whether a real or a virtual source is playing.

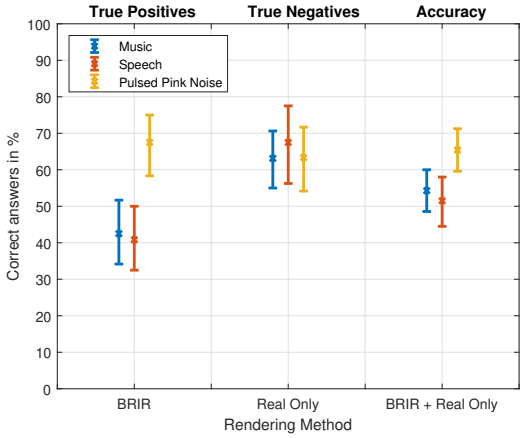
Further, we separately consider the percentage of trials in which participants correctly recognized a scene in which only real sources were active, i.e., true negative



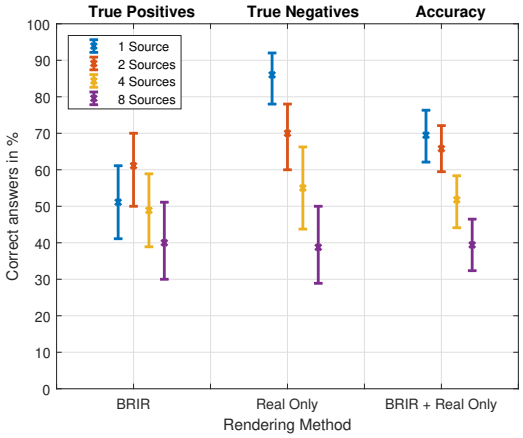
(a) Ratings for all stimuli - HRTF Rendering



(b) Ratings for all complexity levels - HRTF Rendering



(c) Ratings for all stimuli - BRIR Rendering



(d) Ratings for all complexity levels - BRIR Rendering

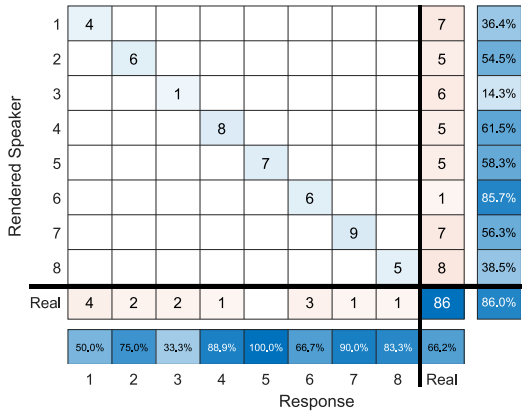
Fig. 4: Mean percentage of correct answers, averaged for all participants. 95% confidence intervals, obtained using bootstrapping. The left part of the plot shows the true positive ratings for HRTF (a), (b) and BRIR (c), (d) rendering, respectively. The middle part shows the true negative rate for only real sources, while the right part shows the total accuracy rating.

in Fig. 4b for the HRTF session and in Fig. 4d for the BRIR session. For real-only rendering, the most apparent effect of the increased complexity is observed. Consequently, the participants are more likely to confuse a real source being a virtual rendering, the more sources are active. Conversely, the participants are less reliable to detect virtualization impairments with higher scene complexity.

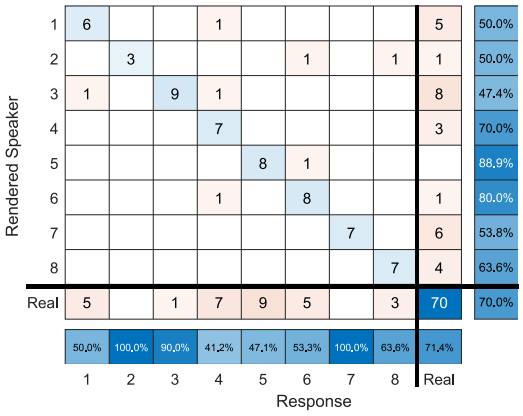
The number of sources that were correctly identified as real depends on the expectation of the listener towards the rendering. We have used the BRIR rendering for the subsequent confusion analysis. The resulting accuracy rating similarly follows the trend to decrease with increasing scene complexity, see right part of the plot in Fig. 4d.

4.3 Classification Errors

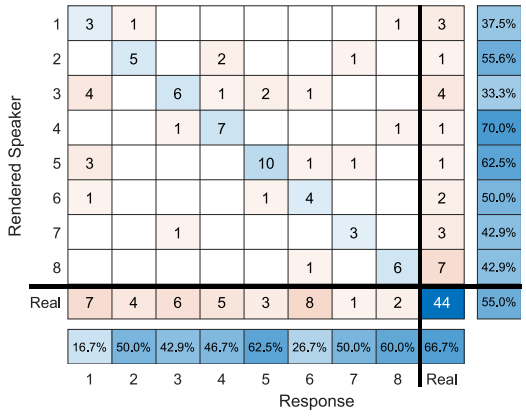
The confusion matrices presented in Fig. 5 provide a more detailed analysis of the responses.In Figure 6, the results from the confusion matrices are summarized as donut plots. The inner ring of the plot corresponds to the overall correct and wrong answers, whereas the outer ring subdivides the ratings regarding virtual or real stimulus and error type. In Figure 6, the answers



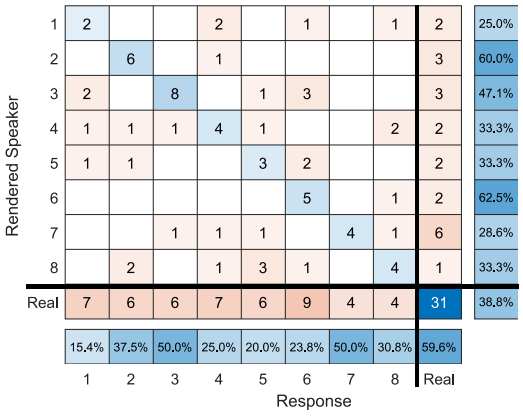
(a) Scene Complexity 1: One sound source



(b) Scene Complexity 2: Two sound sources



(c) Scene Complexity 3: Four sound sources



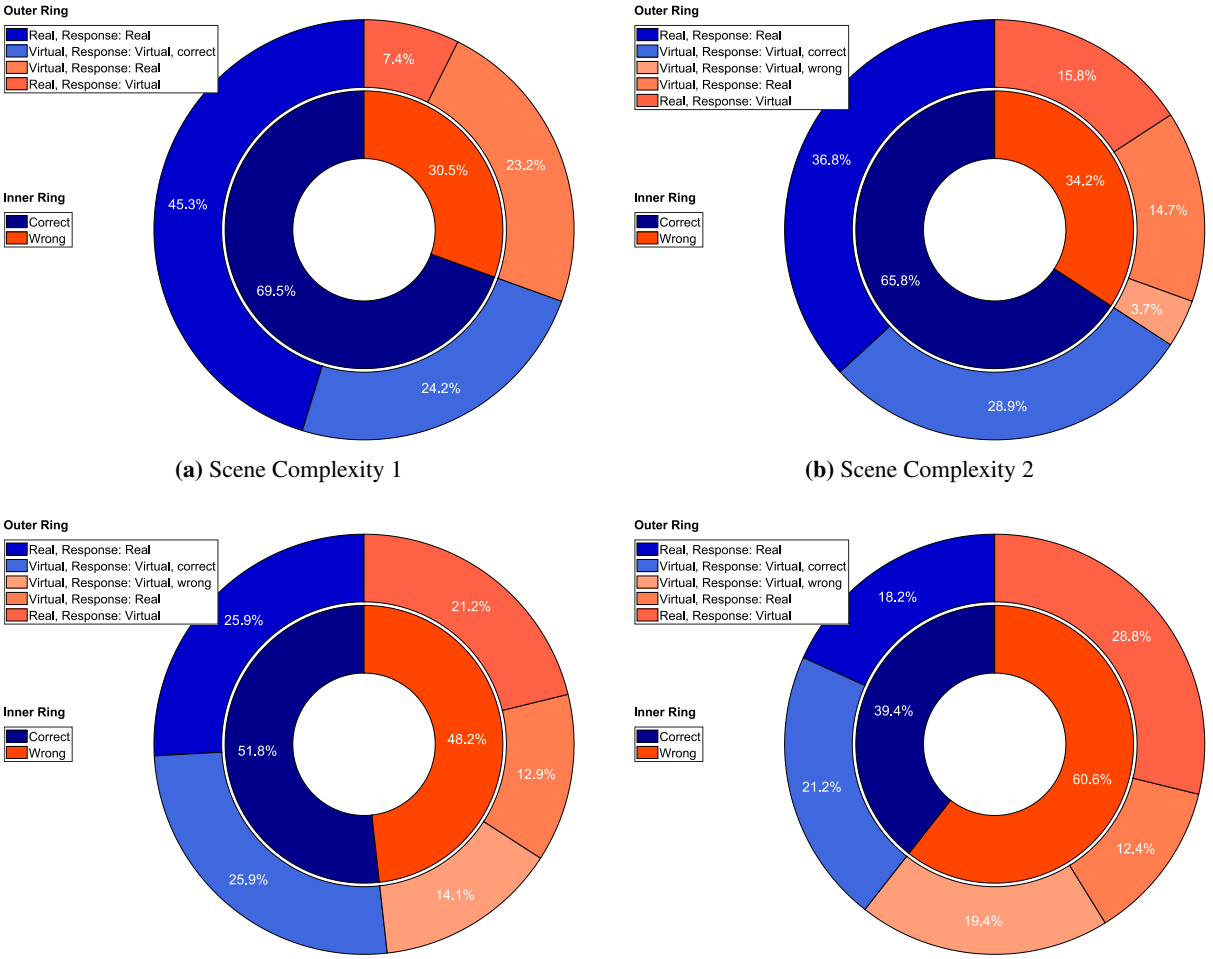
(d) Scene Complexity 4: Eight sound sources

Fig. 5: Confusion matrix of the listening test results for four different scene complexities. In each scenario, there is either none or one sound source is virtual. Matrix entries with a blue shade indicate correct answers, while red shading indicates an incorrect answer. The marginal probabilities are noted on the sides in percentages of correct answers.

are denoted after following rule: (Virtual/Real (Playing Speaker), Response: Virtual/Real (Given Answer, correct/wrong (speaker position of virtualized speaker identified correct/wrong).

In the single source case, the source was able to take the position of all eight speakers. In Fig. 5a, only the first eight diagonal elements in the confusion matrix are non-zero. Thus, all participants were able to localize the source correctly such that errors solely result from confusing real and virtual sources, i.e., non-zero elements in the last row and last column. Therefore, the donut plot in Fig. 6a for the single-source case shows no "Virtual, Response: Virtual, wrong" answers. This proportion increases with complexity. Further, for a single source, a real source was rarely identified as virtual (false positive), but virtual sources were identified as real (false negative) much more commonly. The observed bias towards false negative vanishes in the case of two, four or eight speakers; see Fig 6b, 6c and 6d.

In Figure 6, the donut plots summarizes the confusion matrices. Most importantly, the off-diagonal entries of the first eight rows and columns are collected in the category "Virtual, Response: Virtual, wrong". Notably in Fig. 6d, the "Virtual, Response: Virtual, wrong" are substantially below chance level (39%), which supports



(a) Scene Complexity 1

(b) Scene Complexity 2

(c) Scene Complexity 3

(d) Scene Complexity 4

Fig. 6: Accumulated correct and wrong answers for different scene complexities based on the results shown in Fig 5. The correct guesses depicted in the inner ring are equivalent to the mean accuracy in Fig. 4d. The outer ring shows a more detailed insight of the correct and wrong answers.

that the test implements a simultaneous identification and detection paradigm, where the identification task to choose the correct source position is easier than the detection task of whether there is a virtual source at all.

One effect that is not shown by the presented results, is a learning effect throughout the experiment. Judging from the observations made from the individual results, the participants seem to identify the virtual sources more reliable with ongoing time/trials. Unfortunately, the investigation of this effect was out of the scope in this paper and will be conducted in future research.

5 Summary/Outlook

We have presented a new design idea for experiments in mixed reality audio, which is based on transfer-plausibility, rather than on authenticity or plausibility. This resembles a more realistic application scenario, in which several real-world sound sources are present and no direct comparison between real and virtual sound sources is available. Moreover, it may provide a tool for studying rendering methods of different quality, which would otherwise be impossible due to ceiling effects. Further potential studies might be based on mismatches in room simulation.

The results suggest that when varying scene complexity, the correct identification of virtual sounds is reduced even for low-quality rendering. Also, we have seen that rendering using dynamic, but non-individualized HRTFs and headphone compensation, as well as a static BRIRs, yield surprisingly low detection rates. In this higher quality case, the dependence on scene complexity seems to be smaller.
In the future, the experiment will be conducted with a larger number of participants. Due to the immediate task, this test may be presumably conducted with naive listeners as well. For better comparison of the results between scene complexities and analysis of the individual biases, the real only case will be excluded. Signal detection theory will be applied in order to develop a full theory of transfer-plausibility experiments. Another question concerns the reasons for the decreased accuracy in more complex scenes, and whether it can be explained only by spectral masking.

6 Acknowledgements

This research has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 812719.

The authors want to thank the anonymous reviewers for their detailed and helpful comments on this paper.

References

[1] Blauert, J., *Spatial Hearing*, p. 358, MIT Press, Cambridge, Massachusetts, 1983.

[2] Brinkmann, F., Lindau, A., and Weinzierl, S., "On the authenticity of individual dynamic binaural synthesis," *The Journal of the Acoustical Society of America*, 142(4), pp. 1784–1795, 2017.

[3] Oberem, J., Masiero, B., and Fels, J., "Experiments on authenticity and plausibility of binaural reproduction via headphones employing different recording methods," *Applied Acoustics*, 114, pp. 71–78, 2016.

[4] Langendijk, E. H. A. and Bronkhorst, A. W., "Fidelity of three-dimensional-sound reproduction using a virtual auditory display," *The Journal of the Acoustical Society of America*, 107(1), pp. 528–537, 2000.

[5] Kuhn-Rahloff, C., *Realitätstreue, Natürlichkeit, Plausibilität*, Springer Heidelberg, 2012.

[6] Neidhardt, A., Tommy, A. I., and Pereppadan, A. D., "Plausibility of an interactive approaching motion towards a virtual sound source based on simplied BRIR sets," in *144th Convention*, 2018.

[7] Enge, K., Frank, M., and Höldrich, R., "Listening experiment on the plausibility of acoustic modeling in virtual reality," in *Fortschritte der Akustik - DAGA*, 2020, submitted.

[8] Lindau, A. and Weinzierl, S., "Assessing the Plausibility of Virtual Acoustic Environments," *Acta Acustica united with Acustica*, 98(5), pp. 804–810, 2012.

[9] Pike, C., Melchior, F., and Tew, T., "Assessing the Plausibility of Non-Individualised Dynamic Binaural Synthesis in a Small Room," in *55th International AES Conference*, 2014.

[10] Pulkki, V. and Karjalainen, M., *Communication Acoustics - An Introduction to Speech, Audio and Psychoacoustics*, Wiley, 2015.

[11] Begault, D. R., Wenzel, E. M., and Anderson, M. R., "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *Journal of the Audio Engineering Society*, 49(10), pp. 904–916, 2001.

[12] Völk, F., "Externalization in data-based binaural synthesis: effects of impulse response length," in *Proc. Intern. Conf. on Acoustics (NAG/Tagungsband Fortschritte der Akustik-DAGA 2009), Rotterdam, The Netherlands*, pp. 1075–1078, 2009.

[13] Brimijoin, W. O., Boyd, A. W., and Akeroyd, M. A., "The contribution of head movement to the externalization and internalization of sounds," *PloS one*, 8(12), 2013.

[14] Hendrickx, E., Stitt, P., Messonnier, J.-C., Lyzwa, J.-M., Katz, B. F., and De Boishéraud, C., "Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis," *The Journal of the Acoustical Society of America*, 141(3), pp. 2011–2023, 2017.

[15] McCormack, L. and Politis, A., "SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods," in *AES International Conference on Immersive and Interactive Audio*, 2019.

[16] Niehorster, D. C., Li, L., and Lappe, M., "The accuracy and precision of position and orientation tracking in the HTC vive virtual reality system for scientific research," *i-Perception*, 8(3), 2017.

[17] Borges, M., Symington, A., Coltin, B., Smith, T., and Ventura, R., "HTC Vive: analysis and accuracy improvement," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2610–2615, IEEE, 2018.

[18] Bomhardt, R., Berzborn, M., Klein, J., Richter, J.-G., and Vorlaender, M., "The ITA-Toolbox: An Open Source MATLAB Toolbox for Acoustic Measurements and Signal Processing," 2017.

[19] Zahorik, P., Brungart, D. S., and Bronkhorst, A. W., "Auditory distance perception in humans: A summary of past and present research," *ACTA Acustica united with Acustica*, 91(3), pp. 409–420, 2005.

[20] Harker, A. and Tremblay, P. A., "The HISSTools Impulse Response Toolbox: Convolution for the Masses," in M. Marolt, M. Kaltenbrunner, and M. Ciglar, editors, *ICMC 2012: Non-cochlear Sound*, pp. 148–155, The International Computer Music Association, 2012.

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

[22] Masiero, B. and Fels, J., "Perceptually robust headphone equalization for binaural reproduction," 2012.

[23] Gardner, W. G., "Efficient convolution without input/output delay," in *Audio Engineering Society Convention 97*, Audio Engineering Society, 1994.

[24] Leckschat, D., Epe, C., Dahlheimer, N., Bier, M., Prinz, L., and Schulte, H., "Tonproduktion zweier Titel des Funk-Ensembles "Rhythmussportgruppe"," 2020.

[25] Macmillan, N. A. and Creelman, C., *Detection Theory - A user's guide*, Lawrance Erlbaum Associates, Mahwah, New Jersey, 2nd edition, 2005.