Transfer-Plausibility of Binaural Rendering with Different Real-World References

Nils Meyer-Kahlen[1], Sebastià Amengual Garí[3], Thomas McKenzie[1]
Sebastian J. Schlecht [1,2], Tapio Lokki[1]

Email: nils.meyer-kahlen@aalto.fi

[1] *Acoustics Lab, Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland*

[2] *Media Lab, Department of Art and Media, Aalto University, Espoo, Finland*

[3] *Reality Labs Research, Meta, 8747 Willows Road, Redmond, Washington 98052, USA*

Introduction

The evaluation of virtual acoustics in extended realities (XR), where real and virtual sound sources may co-exist, can be performed using three paradigms: *authenticity*, *plausibility* and *transfer-plausibility*, see Fig. 2. In this paper, we first revisit these three paradigms, before describing a transfer-plausibility experiment, in which participants are asked to identify a virtual source amongst different real sources.

Under the *authenticity* paradigm, the aim is to implement rendering in which a real source and a virtual version thereof cannot be distinguished under all circumstances. To assess such perceptual indistinguishability, discrimination tasks between real sound sources and virtual renderings presented over headphones are performed [1]; suitable experimental designs are forced choice paradigms like ABX tests. Note that in the case of authenticity, it is not required for participants to correctly assign which representation is real and which is virtual, but any audible difference would contradict indistinguishability and thereby also authenticity. Authenticity represents a very strict requirement that is hard to meet in practice, as the just noticeable differences (JNDs) of spectral changes are easily within the range of errors caused by non-individualized binaural playback, or headphone replacement variability. Fortunately, direct comparisons between a real source and a virtual version thereof is not possible in practical XR applications and so authenticity is not required.

When the aim is *plausibility*, direct comparisons between real and virtual sources are not conducted. Here, the question is instead if a virtual sound source presented in isolation is believed to be real. A synonym for plausibility in this sense would be *believability*. Note that this precise sense differs from some uses of the word plausibility, as an attribute more akin to *naturalness*, which can be rated on a continuous scale and does not require confusion between virtual and real sources. In case of plausibility experiments, listeners need to rely only on their expectation of a sound in the real world, sometimes called an *inner reference* [2], when listening to the rendering, without the possibility to compare it to an explicit real reference [3].

The concept of plausibility is most applicable to VR applications. However, for augmented reality (AR) applications, it can be expected that different real sound sources

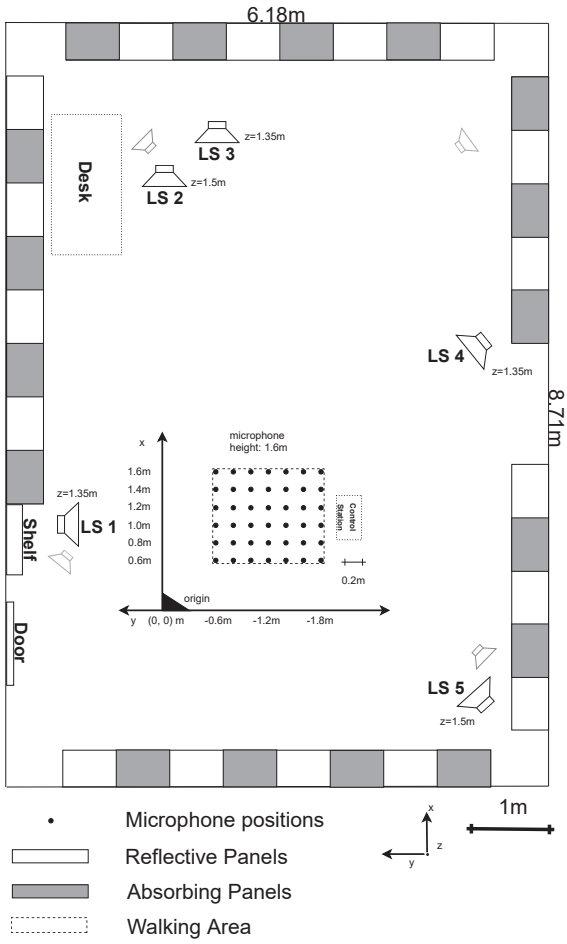


Figure 1: The loudspeaker setup in the variable acoustics room "Arni". Five Genelec 8331AP (black) were used for real sources in the test and four Genelec 8010 (gray) were used for background babble noise in some conditions.

are active alongside virtual sources, emitting different signals from different spatial locations, leading to a certain degree of comparability between real and virtual sound. For this, we refer to the notion of *transfer-plausibility* [4]. Even though plausible or transfer-plausible rendering seems more easily achievable than authentic rendering, creating a stable auditory illusion of this kind in all situations still remains a technically challenging problem.

Here, we present experiments used to assess different degrees of controlled comparison between real and virtual sources on the transfer-plausibility of speech and music in 5 Degrees-of-Freedom (5DoF) rendering environment (movement in the x-y-plane and head rotation).

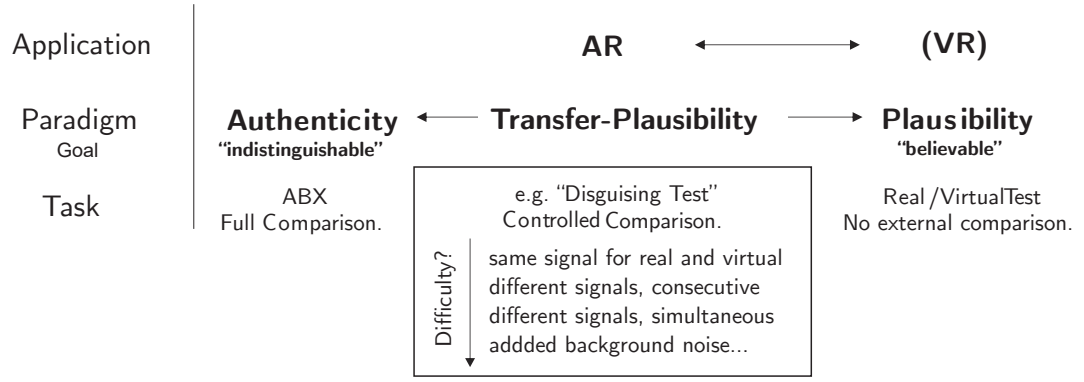


Figure 2: Comparison of the paradigms of Authenticity, Plausibility and Transfer-Plausibility, with tasks that are suitable for assessing them.

The test design was a *disguising test*, in which one virtual source was hidden amongst a varying number of consecutive or simultaneously presented real sources. Here, we want to make a distinction between simultaneous and consecutive real sounds in transfer-plausibility evaluation. Comparing results from past experiments suggests that a large difference may exist: while in [4], adding real sources made detecting the virtual source harder, in [5] adding a real source to the overall experiment lead to more correct identifications. The crucial difference is that while in the former, several sources were added simultaneously, in the latter, the influence was only due to including a real source in the test, so that it was not active simultaneously. The headphone rendering system used in the present study is measurement-based, and can be seen as a baseline, similar to the 3DoF+ implementation used in [4].

Measurements and Rendering

Spatial room impulse responses were measured using the MH Acoustics eigenmike em32 spherical microphone array and five Genelec 8331AP loudspeakers, by means of 2 s long exponential sine sweeps. The variable acoustics room "Arni" at Aalto University, Finland was set to 50% absorptive, leading to a reverberation time of $T_{30} = 0.55$ s. The measurement positions were arranged along a cartesian grid with 20 cm resolution, such that 70 measurements were made between -0.6 m $\leq y \leq 1.2$ m and 0.8 m $\leq x \leq 1.6$ m. The microphone was placed at an approximate ear height of $z = 1.6$ m, see Fig. 1. Precise placement of the microphone was ensured by tracking it using an Optitrack tracking system with six Prime 13 W cameras, which were also used in the subsequent listening experiment.

Even though the measurements and the used rendering software would warrant linear Ambisonics rendering with fourth-order for the entire response, this test used a simpler rendering system, more similar to the rendering used in [4], but inherently incorporating source directivity and source distance. First, the direct sound was extracted from the measurement. Then, for each point of a dense 5×5 cm interpolation grid, the magnitude of the direct sound frequency response of the four closest measurements was weighted with the inverse distance between measurement and interpolation point, summed, and normalized according to the linear sum of weights. Then, a minimum phase pulse was created, the expected direction of the direct sound seen from the interpolation point was calculated, and the direct sound pulse was encoded to fourth-order Ambisonics. All direct sound responses were then saved as a sofa file and loaded into a new VST plugin, created using modules from the spatial audio framework (SAF), called `6DoFconv`[1]. On each frame, the plugin selects the nearest SRIR from the interpolated grid. It then performs convolution, and then the Ambisonics signal is rotated according to the listener's orientation. Decoding was performed using a the magnitude least squares decoder [6]. An earlier version was used and described in [7]. For the present test, the plugin was hosted in Cockos Reaper, whereas the experimental interface was designed in Pure Data, which communicates with Reaper through OSC messages. Head tracking was implemented using an Optitrack system, which sends the tracking data to MATLAB that in turn sends the positions directly to `6DoFconv` via OSC. As a headphone, the Mysphere 3.2 was used, which is designed to be highly transparent to real world sound.

The rendered room reverberation was, as in [4] a single static binaural room impulse response (BRIR) measurement, adjusted in level to the rendered direct sound, in order to achieve the correct direct-to-reverberant ratio. Also as in [4], the BRIR was measured using Sound Professionals TFB-2 binaural microphones. We are well aware that using a single static BRIR is highly nonphysical, however, it provided good results in [4], and therefore we were interested in to see whether allowing listener movement would make this simple solution less viable.

There are several components in the signal processing chain that introduce coloration, such as the measurement microphone or the headphone reproduction. Instead of equalizing individual components, overall equalization was performed by measuring a BRIR of one loudspeaker response and one rendered response. Then, a minimum phase filter was designed to adjust the direct sound and the reverberant tail to the reference BRIR.

As test signals, different speech and music samples from the EBU SQUAM CD, the TSP speech database and the

[1] `https://leomccormack.github.io/sparta-site/docs/plugins/sparta-suite/#6dofconv`

Num. Real Sources	Presentation	Background Noise	Signal
1-4	Consecutive	None	Speech
1-4	Simultaneous	None	Speech
1-4	Consecutive	Babble	Speech
1-4	Simultaneous	Babble	Speech
1-4	Consecutive	None	Music
1-4	Simultaneous	None	Music
1-4	Consecutive	Babble	Music
1-4	Simultaneous	Babble	Music

Table 1: Overview of the experimental conditions.

anechoic recordings available from openairlib [2] were used to assure that participants would not get accustomed to particular samples throughout the test. After all, the objective was to compare consecutive and simultaneous presentation within trials, so comparison between trials should be kept low in this way. Twenty seconds long segments with a random starting time were extracted from the signals, which were then adjusted to have the same root mean squared (RMS) value. The level of one real and virtual source was adjusted to an average of approximately $L_{\mathrm{A}} = 68$ dB at the ear canal entrance of a listener.

In some experimental conditions, diffuse background noise was added, which was generated using a recently developed 'babble noise' generation technique [8], based on phase modification. Four uncorrelated realizations were created and played through the four Genelec 8010 loudspeakers surrounding the listener, see Fig. 1. The noise level was also set to $L_{\mathrm{A}} = 68$ dB at the listeners ear, so that the signal-to-noise ratio for one signal is effectively 0 dB in the cases were background noise was added.

Experimental Design

In each trial of the experiment, there was one virtual source and between one and four real sources. The sources were either presented consecutively, with 5 s duration each, or simultaneously for 20 s. For each of these cases, there was one trial with speech sources, and one trial with musical sources. All the trials were conducted once with no background noise, and once with babble noise. An overview of the conditions is given in Table 1. The order of trials was randomized for each participant.

The experimental GUI showed one tickbox for each of the five loudspeakers and participants were asked to select which of the sources was virtual. Further, the interface gave the opportunity to "enter the main cue you used, and possible comments". Participants were instructed to provide the reason for their choice in this box, especially if they were certain of their answer. Further, participants were instructed to move freely in the walking zone (see dotted lines in Figure 1), but not to rotate too abruptly, to ensure the headphone would not move too much or even fall off during the test. Also, the instruction clarified that the main focus should not be on testing the quality of the tracking.

[2] https://openairlib.net/

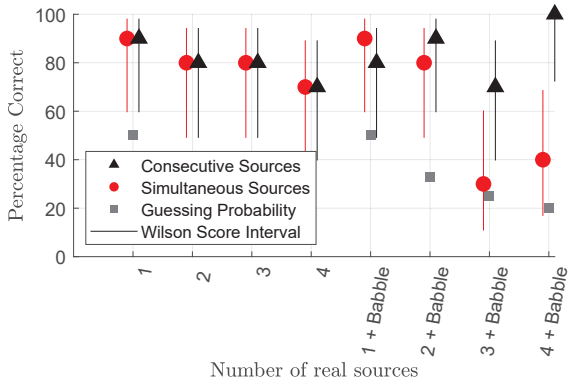
Results and Discussion

Ten participants took part in the test (age M = 25.6 y, STD = 2.4 y); five were members of the Aalto Acoustics Lab, experienced with listening tests, and five were Masters students with less experience. The experience level was assessed by asking about the years spent with acoustics research, audio engineering and professional musical training, and a subjective experience rating in the same fields. Interestingly, correlating the participants experience data to the identification performance did not reveal any noteworthy patterns.

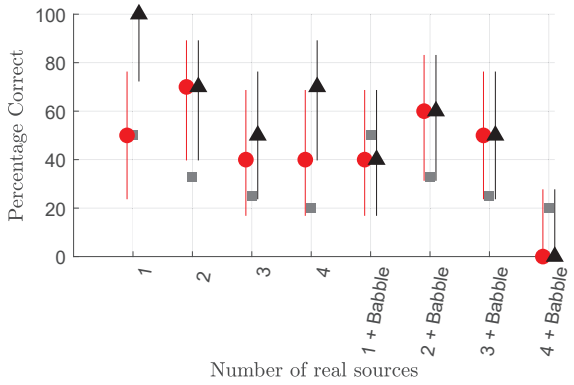
When analyzing the percentage of correct answers shown in Figure 3, the first apparent result is the relatively high number of correct identifications for the speech stimuli. It indicates that for speech, participants often identified the virtual source correctly, so the system cannot be said to accomplish transfer-plausible rendering in all scenarios. It is interesting that for both presentation modes, performance dropped with an increasing number of sources. A possible explanation is that even though recognizing a virtual source is possible, remembering and assigning it correctly becomes more difficult. In the very complex scenes comprising three or four different speakers and babble noise, the percentage of correct answers drops severely in the case of simultaneous sources. It appears that here, focusing on individual sources is the most complicated, and virtual sources did not stand out so much as to attract attention. This suggests that virtual sources blend in to a complex scene sufficiently well, even with speech. In contrast, purely the masking provided by the noise to single speech sources in consecutive presentation was not sufficient to reduce identification.

For music, the identification rates were generally lower. In the babble noise-free case, some differences between consecutive and simultaneous presentation were observed. Interestingly, the number of confusions were exactly the same for simultaneous and consecutive presentations, when the babble noise was added, and for the case of four added real sources, no participant gave the correct answer in any case. For the other cases, the character of the babble noise and of the musical signals seemed to be sufficiently different, such that adding it does not influence the ability to focus on certain sources.

The cues that participants stated when they were able to identify the virtual source as regarded as an equally important result, see Table 2. Cues were given in free text, and coding to the attributes in table 2 was done manually by the authors. As expected for a disguising test, in which no direct comparison to a real version of the virtual source is possible, loudness and coloration cues were secondary. Two of the most common responses were *elevation*, and *horizontal localization*. These cues point to typical problems of non-individualized binaural rendering. As it has been recently shown that individualization of head related transfer functions might not play an important role in multimodal, dynamic scenarios [9], this observation requires further analysis. Although not directly apparent, these issues may be related to the static BRIR rendering, which also will be tested in the future.



(a) Speech



(b) Music

Figure 3: Percentage of correct identifications

It is interesting that a lack of *externalization* or in-head localization were rarely reported, which is another common problem of such rendering. However, *proximity* was commonly mentioned.

Conclusion

This paper described a disguising test under the paradigm of transfer-plausibility, an experimental design for AR audio evaluation. Participants had to identify a virtual sound source, presented amongst a varying number of either consecutive of simultaneous real sources. For speech, identification rates were high in many conditions, indicating that the used baseline rendering system did not create transfer-plausible rendering under all circumstances. Only for very complex scenes, comprising three or four virtual sources and background noise, identification dropped. For music however, identification was overall lower. Furthermore, with the addition of background babble, consecutive and simultaneous representation were the same for musical sources.

In the future, the system will be extended to provide the infrastructure for even more transfer-plausibility experiments, also including visual presentation using a head mounted display. Also, as the main cues used for identification related to generic binaural rendering, the importance of different perceptual cues on transfer-plausibility will be assessed in upcoming tests in more detail.

Cue	Total Mentions	Participants mentioned
Elevation	33	6
Extra sound from above	13	1
Proximity	16	3
"Air"/Hi. Freq. too close	2	2
Horizontal localization	14	6
Externalization	4	1
Front-Back confusions	2	2
Width	4	4
Localizeability	4	2
Jumping / Dropouts	7	3
Loudness	3	3
Frequeny response	4	2
Sibilants too strong	1	1

Table 2: Cues mentioned by the participants.

Acknowledgments

This research was supported by the EU's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 812719 and Human Optimised XR (HumOR) Project, funded by Business Finland.

References

[1] Brinkmann, F., Lindau, A., and Weinzierl, S.: On the authenticity of individual dynamic binaural synthesis. J. Acoust. Soc. Am. 142 (2017), 1784–1795

[2] Kuhn-Rahloff, C.: Prozesse der Plausibilitätsbeurteilung am Beispiel ausgewählter elektroakustischer Wiedergabesituationen. PhD thesis, TU Berlin, 2011

[3] Lindau, A. and Weinzierl, S.: Assessing the Plausibility of Virtual Acoustic Environments. Acta Acust united Ac. 98 (2012), 804–810

[4] Wirler, S. A., Meyer-Kahlen, N., and Schlecht, S.J.: Towards Transfer-Plausibility for Evaluating Mixed Reality Audio in Complex Scenes, AES AVAR (2020)

[5] Neidhardt, A., Zerlik, A. M.: The Availability of a Hidden Real Reference Affects the Plausibility of Position-Dynamic Auditory AR. Front. Virtual Reality 2 (2021), 678875

[6] Schörkhuber, C., Zaunschirm, M., Holdrich., R.: Binaural Rendering of Ambisonic Signals via Magnitude Least Squares. DAGA (2018)

[7] McKenzie, T., Schlecht, S.J., Pulkki, V.: Auralisation of the Transition between Coupled Rooms. I3DA (2021)

[8] Kuusinen, A.: Phase multiplied babble noise for speech-in-noise tests (2022), in review

[9] Rummukainen, O. S., Robotham, T., and Habets, E. A. P.: Head-Related Transfer Functions for Dynamic Listeners in Virtual Reality', Appl. Sci. vol. 11, no. 14, (2021), 6646