The role of source signal similarity in distinguishing between different positions in a room

Audio Engineering Society

Conference Paper

Presented at the International Conference on
Spatial and Immersive Audio
2023 August 23–25, Huddersfield, UK

The role of source signal similarity in distinguishing between different positions in a room

Thomas McKenzie[1], Nils Meyer-Kahlen[2], and Sebastian J. Schlecht[2,3]

[1]*Acoustics and Audio Group, Reid School of Music, University of Edinburgh, United Kingdom*
[2]*Acoustics Lab, Department of Information and Communications Engineering, Aalto University, Espoo, Finland*
[3]*Media Lab, Department of Art and Media, Aalto University, Espoo, Finland*

Correspondence should be addressed to Thomas McKenzie (thomas.mckenzie@ed.ac.uk)

ABSTRACT

Typically, evaluation of spatial audio systems uses the same source signal for each condition in listening comparison tests (such as ABX and MUSHRA). However in an augmented reality scenario, it is unlikely that the exact same source signal would exist at the exact same position in space, both real and virtual: instead, a real source would be in one position in the room and a virtual source in a different position, both with different source signals. A perceptual study is presented on the effect of source signal similarity when distinguishing different positions in a room. Three source signal types (all speech) are investigated in a multiple stimulus paradigm: the same source signal for all conditions, the same speaker but a different sentence for each condition, and a different speaker and different sentence for each condition. Results show that the source signal similarity significantly impacts the similarity rating between different receiver positions in the same room, which suggests that spatial audio system fidelity requirements could vary depending on the source signal types used in the target application.

1 Introduction

Evaluation of virtual acoustics for augmented and virtual reality (AR/VR) is faced with a challenge. On the one hand, renderings of virtual sound sources often appear perfectly natural and free of artefacts when listened to in isolation. On the other hand, when the same virtual source can directly be compared to a reference, e.g., a binaural recording or the same sound source reproduced in the real room, even high quality algorithms are often audibly different. These differences result from physical limitations of employed microphone arrays [1], trade-offs due to computational limitations when rendering in real time including latency [2, 3], lack of HRTF individualisation and headphone equalisation [4], or inaccurate properties in room acoustic simulations [5].

Therefore, researchers and developers have to ask themselves if virtual acoustics should strive for *authenticity* [6], i.e., the exact perceptual indistinguishability between a rendering and a real reference, or if they should find other ways of evaluation, acknowledging that direct comparisons are not possible in practise. In AR for example, a reference is only given through other sound sources that may be present in the real room, emitting different signals. For instance, in a mixed reality teleconference scenario, people listen to one speaker in the real room and another speaker who is virtual.

Alternative evaluation paradigms without an authentic direct reference are being explored, such as *plausibility* [7, 8]; assessing whether a rendering is believed to be real when it can only be compared to one's inner expectations, or *transfer-plausibility* [9, 10]; comparing real and virtual, but using different sound sources in different locations in the same room. Currently, work is focused towards better defining these paradigms, to make experiments more comparable [11]. However, AR/VR evaluation requires real-time rendering, and there is an inherent risk of many confounding factors. For example, if listeners can move freely to test the rendered sound, there will be variability in which parts of the scene are explored [12, 13, 14].

Due to the practical challenges of such tests and possibly also due to tradition (and researchers' habits), it is still common to test specific AR/VR algorithms under better known and tested multiple stimulus comparison, MUSHRA-like paradigms. This paper postulates that differences may exist which are only perceived in such unrealistic tests allowing for comparison to a reference rendered with the same source signal as those used for other conditions, and that there is a risk for making false conclusions about the relevance of specific rendering aspects for sound for AR/VR. This would be in keeping with a recent study on concert hall recognition [15], in which listeners were able to correctly identify a concert hall from three others using the same musical excerpt, but were much less accurate when each option used a different musical excerpt.

To illuminate the issue, this paper proposes to study the role of source signal similarity in a room acoustic experiment. The tested room acoustic conditions are hybrid binaural room impulse responses (BRIRs), where the first part of the response is taken from a measurement made at one position in the room, and the second part is taken from a different measurement made at a different position in the room. The design is based on a test conducted recently in [16]. The term 'source signal' is used in this paper to describe the stimulus to be convolved with the room acoustic characteristics that make up the conditions of a multiple stimulus test. The 'authentic' reference is used to refer to a hidden reference condition that uses the same source signal as the open reference condition (as is traditionally done), and 'transfer' reference is used to refer to a hidden reference condition made using a different source signal, but convolved with the same transfer function as the open reference.

First, the results of a regular MUSHRA test are shown, and then the test is re-run with the same room acoustic conditions, but rendered using different speech excerpts. A comparison is made between the standard approach using the same source signal (one sentence spoken by one speaker), renderings with different sentences spoken by the same speaker, and finally different sentences spoken by different speakers. In the regular MUSHRA test, the hidden reference is the 'authentic' reference. In the two tests with different source signals for each condition, there is both an 'authentic' hidden reference and a 'transfer' reference.

The paper is laid out as follows: Section 2 introduces the test methodology in detail, Section 3 shows the results of the experiment, which are discussed in Section 4, along with suggestions for improvements and further experiments. Summary and conclusions are found in Section 5.

2 Methods

To assess the role of source signal similarity in room position perception, a listening test was conducted under a multiple stimulus comparison paradigm. There were three variables that made up the test. Firstly, three types of base source signal (to be convolved with the test measurements for each condition) were used, ranging from entirely the same signal to significantly different. Secondly, for each source signal type under test, three trials were conducted; these were for different positional changes in the room, to provide variety in the tested room acoustics scenarios. Finally, in each trial, five versions of the positional changes were tested, where the switching time between the first part (stemming from BRIR1) and the second part (from BRIR2) of the response were varied.

2.1 Source Signals

The three different source signal selection schemes for the base source signals used in the test are as follows:

1. *Same speaker, same sentence*: One speaker reciting one sentence, used for all conditions.

2. *Same speaker, different sentence*: The same speaker, but reciting a different sentence for each condition.

3. *Different speaker, different sentence*: A different speaker reciting a different sentence for each condition.

The sentences spoken were phonetically balanced 'Harvard' sentences [17]. The *same speaker* recordings were anechoic recordings of a male speaker [18], and the different speaker recordings were three male, four female speakers [19]. All recordings used a 48 kHz sampling rate, and each excerpt was normalised to the same root-mean-square (RMS) amplitude.

2.2 Positions in the Room

Three different combinations of source and receiver positions in a room were used, which followed those used in [16]; a recent study that included a listening test on room acoustics similarity. The three pairs of room acoustics measurements correspond to the source – receiver (S–R) positions as follows (and illustrated in Fig. 1):

1. BRIR1: S1–R7 ; BRIR2: S1–R2.
2. BRIR1: S2–R5 ; BRIR2: S2–R2.
3. BRIR1: S3–R4 ; BRIR2: S3–R1.

The BRIRs were rendered from 3rd order Ambisonic spatial room impulse responses (SRIRs) from the variable acoustics 6DoF dataset[1] (as described in [20]). The measurements used here were the most reverberant, with octave-band [500 Hz, 1 kHz, 2 kHz] RT$_{60}$ values of [1.28, 122, 1.12] s, recorded at a background noise level of 20.5 dB SPL(A). The binaural rendering of the SRIRs used dual-band Ambisonic decoding (mode-matching below 1.8 kHz, max-$\mathbf{r}_\mathrm{E}$ above), with time-alignment above 1.8 kHz [21] and Ambisonic diffuse-field equalisation [22] pre-processing[2], and non-individualised Neumann KU100 HRTFs [23]. Rendered BRIRs were then time-aligned to the direct sound exceeding a threshold of 1/10 the maximum absolute pressure. The binaural rendering was chosen to be static (without head-tracking) to ensure a consistent and controlled listening experience between participants.

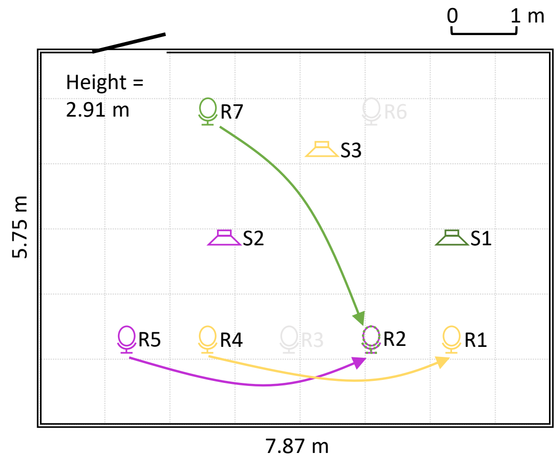


Fig. 1: Illustration of the source (S) and receiver (R) positions used to make up the listening test conditions. Microphones and loudspeakers were oriented faced north and south, respectively, according to the illustration orientation. Note that R2 was used both for the first and second room position changes.

2.3 Test Conditions

In each trial, there were five test conditions corresponding to different room acoustic conditions. The reference conditions (*Ref A* and *Ref T*) used the BRIR1s S1–R7, S2–R5 and S3–R4. The fifth conditions, herein labelled as *0ms*, used the BRIR2s corresponding to S1–R2, S2–R2 and S3–R1. The receiver changes from the *Ref* to *0ms* conditions are denoted by an arrow in Fig. 1.

The second, third and fourth conditions were hybrids of the reference and fifth conditions, for which the early part was of the *Ref* BRIR1s, and the tail was from the *0ms* BRIR2s. The switching times (from BRIR1 to BRIR2) were 3 ms, 8 ms and 13 ms, which corresponded approximately to the direct sound, the direct sound and first early reflection, and the direct sound and first few early reflections, respectively. This method was chosen to emulate different levels of reproduction accuracy, whereby a later switching time would mean more of the test BRIR was made up of the reference BRIR, producing a more accurate reproduction. No crossfades were applied when switching between BRIRs, and no rotations or other adjustments such as to the direct-to-reverberant ratio were made.

[1] `https://doi.org/10.5281/zenodo.5720724`

[2] `https://github.com/thomas-mckenzie/binaural_ambisonic_preprocessing`

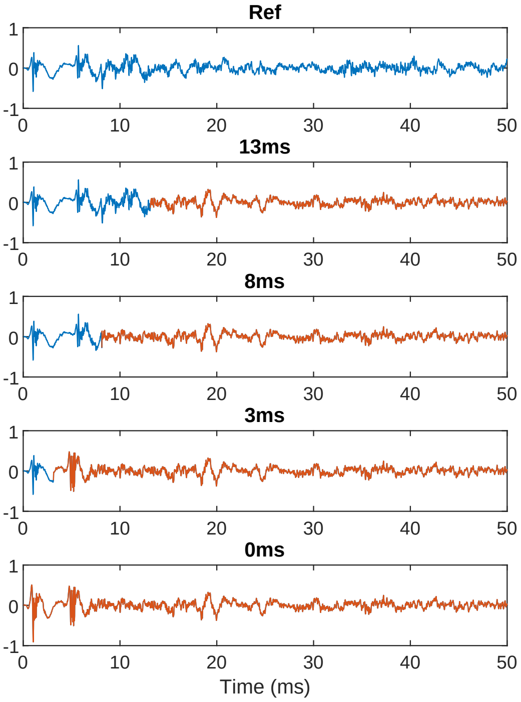


Fig. 2: Conditions showing the switching from BRIR1 to BRIR2 (S2−R5 in blue, S2−R2 in red, as coloured magenta in Fig. 1), left signal.

The five test conditions are as follows (time-domain plots of the first 50 ms of the hybrid BRIRs are shown in Fig. 2):

1. *Ref*: Entire duration is BRIR1.
2. *13ms*: First 13ms is BRIR1, the rest is BRIR2.
3. *8ms*: First 8ms is BRIR1, the rest is BRIR2.
4. *3ms*: First 3ms is BRIR1, the rest is BRIR2.
5. *0ms*: Entire duration is BRIR2.

2.4 Test Paradigm

For the *same speaker, different sentence* and *different speaker, different sentence* stimuli types, two reference conditions were included: the 'transfer' reference, a hidden reference that used the same BRIR as the reference but a different source signal, (labelled in plots as *Ref T*) and the standard 'authentic' hidden reference with the exact same signal as the reference (herein referred to as *Ref A*). The monophonic source signal without convolution was included as an anchor. Note that this differs from the ITU-R BS.1534-3 standard for the MUSHRA test paradigm [24], as the hidden reference would traditionally be the exact same signal (i.e. only the 'authentic' reference) and the anchor would be a low-passed version of the reference.

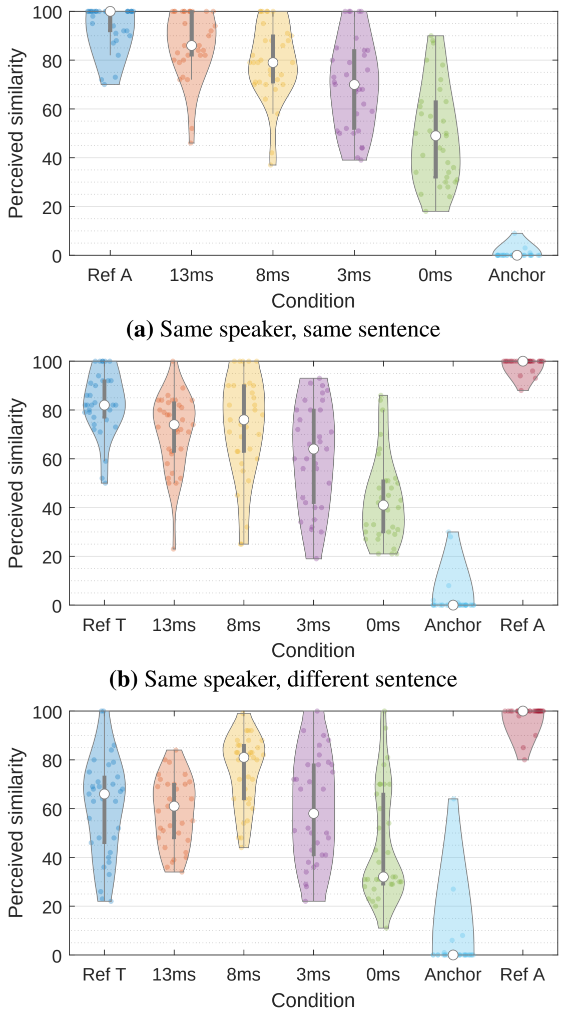
In each trial, participants were asked to rate the similarity of each test condition's acoustic characteristics to those of the open reference condition. They were instructed to consider the sound location, reverberation, distance, and colouration. Participants were explicitly given the example that, if two sounds have different speech and speakers but the same positional, room acoustic characteristics, they should be rated the same. Participants were instructed to use the full scale for each trial, and looping was enabled. Each trial had six or seven conditions (for the same source signal case and the different source signal cases, respectively), which pertained to the different positions in the room plus the anchor.

The test consisted of a total of nine trials: the three source signal types repeated at each of the three room position changes. No trials were repeated. A training trial was included prior to the test beginning, which included *Ref T*, *8ms*, *0ms* and *Anchor* conditions, to ensure participants were familiar with the test software and the task at hand. The playback volume was adjustable during the training phase, after which it was fixed. Test trials and conditions were randomised and presented double blind (neither the participant nor the assessor knew what conditions were being presented). Open back AKG K240 Studio headphones were used by all participants. The test was conducted using the webMUSHRA[3] interface (as described in [25]).

3 Results

11 participants, aged between 21 and 27 (9 male, 2 female), completed the test. No reported hearing issues were stated and participants had prior critical listening experience. The results are presented as violin plots in Fig. 3, where the results of the three different room position changes are collated. Violin plots are used here; they augment the box plot by displaying the density trace [26]. The violin width shows the density of data, median values are presented as white points, interquartile ranges are thick grey lines, range between the lower and upper adjacent values are thin grey lines, and individual results are coloured points.

[3] https://github.com/audiolabs/webMUSHRA

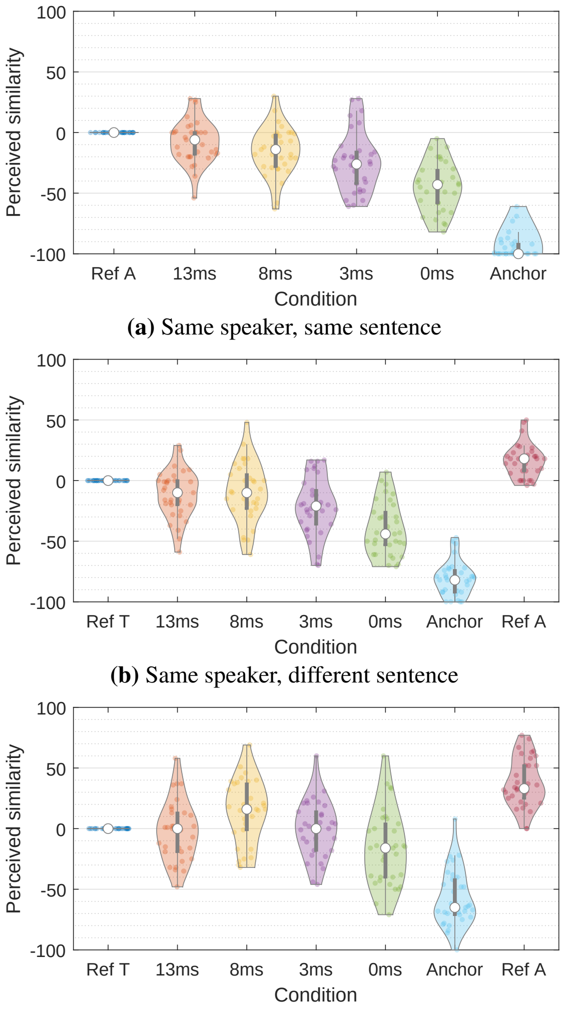


(a) Same speaker, same sentence

(b) Same speaker, different sentence

(c) Different speaker, different sentence

Fig. 3: Violin plots of the listening test results. Median values are a white point, interquartile range a thick grey line, the range between lower and upper adjacent values a thin grey line, and individual results are coloured points. Ref A refers to the 'authentic' reference (same source signal as open reference) and Ref T refers to the 'transfer' reference (different source signal to open reference).



(a) Same speaker, same sentence

(b) Same speaker, different sentence

(c) Different speaker, different sentence

Fig. 4: Violin plots of the listening test results, renormalised relative to the 'transfer' reference score (Ref T). Median values are a white point, interquartile range a thick grey line, the range between lower and upper adjacent values a thin grey line, and individual results are coloured points. Ref A refers to the 'authentic' reference (same source signal as open reference) and Ref T refers to the 'transfer' reference (different source signal to open reference).

Firstly, for the *same speaker, same sentence* source signal, which corresponds to a traditional multiple stimulus comparison, it is clear to see that all conditions were often distinguishable from the reference and that the *0ms* condition is rated the lowest. This is to be expected, as the only difference between test conditions was the position in the room.

For the *same speaker, different sentence* source signal, the trend is largely continued, though the 'transfer' reference with a different source signal (*Ref T*) was rated lower than the 'authentic' reference with the same source signal (*Ref A*). Here, the difference between test conditions was not solely the position in the room, and so participants had to try and remove the source signal's impact and make the comparison. This was still somewhat possible due to the similarities in vocal range, performance and loudness between the source signals, though harder than the *same speaker, same sentence* source signal.

Finally, in the *different speaker, different sentence* source signal, most differences are no longer clearly visible. Even the *3ms* condition, in which effectively only the direct sound is the same as in the reference, does not evoke audible differences. An exception is the *0ms* condition, in which a completely different BRIR is used, so that the direct sound originates from a different direction as well. Surprisingly, the median rating of the *8ms* condition was even higher than the 'transfer' reference *Ref T* (that used a different stimulus). It is clear that here, participants found it difficult to detach the source signal from the position in the room. The large differences in vocal ranges, accents, loudness and performance, had a great effect on the perception of position and reverberation.

As it was not uncommon for the 'transfer' reference to be rated lower than the conditions, especially for the different sentence source signal types, Fig. 4 presents violin plots of the listening test results with a re-normalisation of each condition's result relative to the 'transfer' reference rating of that trial. This shows the relative similarity between the results of the *same speaker, same sentence* and *same speaker, different sentence* source signal types. It also highlights the larger differences of the *different speaker, different sentence* source signal type, where even the *0ms* condition was in some cases rated higher than the reference.

To test the statistical significance of the findings, results data was first tested for normality using the Shapiro-Wilk test. Data was not all normally distributed at a 5% significance level, so non-parametric testing was used. Firstly, a Friedman test comparing the answers for each of the tested source signal types (with results for different conditions collated) was conducted. This was highly statistically significant ($\chi^2(2) = 56.2, p < 0.001$), which shows that the source signal choice has a definite effect on results. Post-hoc pairwise Wilcoxon signed-rank tests, using the Bonferroni-Holm [27] correction, reveal significant differences for the reference and the *13ms* condition. In the case of the reference, the difference between the conventional *same speaker, same sentence* and the *same speaker, different sentence* source signal selection ($p = 0.01$), and between the conventional and the *different speaker, different sentence* ($p < 0.001$) were significant.

Results of pairwise Wilcoxon signed-rank tests with Bonferroni-Holm correction regarding comparison within one source signal selection scheme, are presented in Table 1. Unless stated otherwise, statistical significance is herein declared at $p < 0.05$. Notable results here are that for the *same speaker, same sentence* source signal, all conditions except *13ms* were significantly different from the reference. For the *same speaker, different sentence*, both the *13ms* and *8ms* conditions were not statistically significantly different from the reference. For the *different speaker, different sentence* source signal, however, all conditions except the anchor and *Ref A* (using the same source signal as the reference in the multiple stimulus test) were not statistically significantly different.

4 Discussion

The results show that changing the source signal used for rendering different room impulse responses influences room acoustic similarity ratings. The greatest difference is observed when using speech from different speakers. Using a different sentence spoken by the same person left the inter-condition results mostly unaffected, albeit with lower overall ratings (reference included).

Comparing renderings with a *different speaker, different sentence* to the traditional multiple stimulus test case of using the same source signal for all conditions, it becomes clear that acoustical differences can exist that are only audible in the same source signal case. Significant differences were found for the *13ms* condition. Compare also the *3ms* condition between the source signal selection schemes, where the perceived difference in the *same speaker, same sentence* is large, but seems to disappear in the *different speaker, different sentence* condition.

Possible reasons for this are that different speakers not only have varying registers of voice, but performance variations too. Some people naturally speak louder or softer, enunciate more or less, have different accents and dialects. These significant variations contribute to a

Table 1: Wilcoxon signed-rank tests of the listening test results between the conditions and reference, for the three source signal types (with the Bonferonni-Holm correction). For the top row, Ref refers to Ref A: the 'authentic' reference (same source signal as open reference), whereas for the second and third rows, Ref refers to Ref T: the 'transfer' reference (different source signal as open reference).

Source signal category	Ref	13ms	8ms	3ms	0ms	Anchor	Ref A
Same speaker, same sentence	>1.00	>1.00	**0.01**	**0.003**	**<0.001**	**<0.001**	/
Same speaker, different sentence	>1.00	**0.16**	>1.00	**0.002**	**<0.001**	**<0.001**	**<0.001**
Different speaker, different sentence	>1.00	>1.00	**0.35**	>1.00	**0.27**	**<0.001**	**<0.001**

larger auditory difference than simply altering the room acoustics in which the sound exists. With source signals where some frequencies have differing loudness, this excites and emphasises different features of the room's acoustics.

These result suggest that, due to the lower perceptual sensitivity to room acoustic changes when the stimulus varies significantly from that used for rendering the reference, an augmented reality scenario could potentially use a less complex reverberation rendering method and still produce a convincing perception of the room's acoustic characteristics. Specifically, using the correct direct sound for each source / receiver position, but employing the same BRIR independent of the position (as done in [9, 10]) may be sufficient in some cases, as is shown by the results of the *3ms* condition in Fig 4c. In fact, even the *0ms* condition, using a different direct sound, was not statistically significantly different from the reference at a 95% confidence interval. Using the traditional test design, with the same source signal for all test conditions, one would make a different conclusion. Note that in that case, even the smallest tested difference in room acoustics (the *13ms* condition) was audible.

With regards to the design of the presented experiment, it should be noted that although for each of the 9 trials, the source signal ordering was randomised pre-convolution with the test BRIRs, this was only done once for the test, i.e., every participant listened to the same audio files. This could go part way to explaining how some conditions were surprisingly rated as more similar than the reference in the *different speaker, different sentence* scenario. It is possible that the source signals used in these conditions were closer in timbre and loudness to those used for the corresponding reference conditions. Future iterations of the test can mitigate this by separately randomising the stimulus ordering pre-convolution for each participant.

Future iterations could also use dynamic binaural rendering. It is unknown whether this would make the task of distinguishing different positions in a room easier or harder. One argument against dynamic rendering would be that the addition of head rotations may cause more errors in judgement, due to the complex task of trying to rate different sounds whilst also moving ones head, which is constantly refreshing and changing ones auditory memory, as was found in [14]. However, an argument for dynamic rendering would be that the improved localisation accuracy from the dynamic cues [4] may mean listeners are better able to pinpoint the direction of direct sound and the spatial characteristics induced by the early reflections, and may therefore find the task of distinguishing between different positions in a room easier.

5 Summary

This paper has presented a perceptual study on the effect of source signal similarity when distinguishing different positions in a room, whereby source signal here denotes the stimulus to be convolved with the room acoustic characteristics that make up the conditions of a multiple stimulus test. Three categories of source signal have been tested in a multiple stimulus paradigm for the distinguishing of different positions in a room (all speech): same speaker, same sentence; same speaker, different sentence; and different speaker, different sentence.

The results of this study show that, when the source signal is constant between conditions, it is possible to distinguish between different reverberation conditions effectively. However, when the source signal varies between conditions as well, the task becomes significantly harder. Listeners are less proficient at removing the impact of the source signal and still making the comparison.

These findings have implications for the design and evaluation of systems for rendering virtual environments. In a scenario where real and virtual sound sources use the same source signal, listeners would be able to discern differences in reverberation easily and thus a high accuracy of reproduction would be necessary. Conversely, in a scenario where real and virtual sources use different source signals, and discerning differences in reverberation is more difficult, it may therefore be sufficient to employ a lower accuracy of reproduction.

Samples of the listening test audio files for the three tested source signal types, as well as the non-convolved BRIRs, are available online[4]. This is for the second source-receiver combination, S2−R5 and S2−R2 (source and receiver positions as illustrated in Fig. 1, where the transformation for the different conditions as shown in Fig. 2).

References

[1] Ahrens, J. and Andersson, C., "Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre," *J. Acoust. Soc. Am.*, 145(4), pp. 2783–2794, 2019, doi:10.1121/1.5096164.

[2] Brungart, D. S., Kordik, A. J., and Simpson, B. D., "Effects of headtracker latency in virtual audio displays," *J. Audio Eng. Soc.*, 54(1/2), pp. 32–44, 2006.

[3] Lindau, A., "The perception of system latency in dynamic binaural synthesis," in *Fortschritte der Akustik: Tagungsband der 35. DAGA*, pp. 1063–1066, 2009.

[4] 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," *J. Audio Eng. Soc.*, 49(10), pp. 904–916, 2001.

[5] Brinkmann, F., Aspöck, L., Ackermann, D., Lepa, S., Vorländer, M., and Weinzierl, S., "A round robin on room acoustical simulation and auralization," *J. Acoust. Soc. Am.*, 145(4), pp. 2746–2760, 2019, doi:10.1121/1.5096178.

[6] Brinkmann, F., Lindau, A., and Weinzierl, S., "On the authenticity of individual dynamic binaural synthesis," *J. Acoust. Soc. Am.*, 142(4), pp. 1784–1795, 2017, doi:10.1121/1.5005606.

[7] Lindau, A., Erbes, V., Lepa, S., Maempel, H. J., Brinkman, F., and Weinzierl, S., "A spatial audio quality inventory (SAQI)," *Acta Acustica United with Acustica*, 100(5), pp. 984–994, 2014, doi:10.3813/AAA.918778.

[8] Pike, C., Melchior, F., and Tew, T., "Assessing the plausibility of non-individualised dynamic binaural synthesis in a small room," in *AES Int. Conf. on Spatial Audio*, pp. 1–8, 2014.

[9] Wirler, S., Meyer-Kahlen, N., and Schlecht, S. J., "Towards transfer-plausibility for evaluating mixed reality audio in complex scenes," in *AES Int. Conf. on Audio for Virtual and Augmented Reality*, Online, 2020.

[10] Meyer-Kahlen, N., Amengual Garí, S., McKenzie, T., Schlecht, S. J., and Lokki, T., "Transfer-plausibility of binaural rendering with different real-world references," in *Jahrestagung für Akustik - DAGA 2022*, Stuttgart, 2022.

[11] Quackenbush, S. R. and Herre, J., "MPEG standards for compressed representation of immersive audio," *Proceedings of the IEEE*, 109(9), pp. 1578–1589, 2021, ISSN 15582256, doi:10.1109/JPROC.2021.3075390.

[12] Olli Rummukainen, Robotham, T., Schlecht, S. J., Plinge, A., Herre, J., and Habets, E. A. P., "Audio quality evaluation in virtual reality: multiple stimulus ranking with behavior tracking," in *AES Conf. on Audio for Virtual and Augmented Reality*, Redmond, WA, 2018, doi:10.1016/s0166-5316(99)00015-2.

[13] Robotham, T., Rummukainen, O., Herre, J., and Habets, E. A., "Online vs. offline multiple stimulus audio quality evaluation for virtual reality," in *145th AES Conv.*, New York, 2018.

[4] `http://sebastianjiroschlecht.com/publication/Source-Signal-Similarity/`

[14] McKenzie, T., Meyer-Kahlen, N., Hold, C., Schlecht, S. J., and Pulkki, V., "Auralisation of measured room transitions in virtual reality," *J. Audio Eng. Soc.*, 71(6), pp. 326–337, 2023, doi: 10.17743/jaes.2022.0084.

[15] Kuusinen, A. and Lokki, T., "Recognizing individual concert halls is difficult when listening to the acoustics with different musical passages," *J. Acoust. Soc. Am.*, 148(3), pp. 1380–1390, 2020, ISSN 0001-4966, doi:10.1121/10.0001915.

[16] Deppisch, T., Garí, S. V. A., Calamia, P., and Ahrens, J., "Perceptual evaluation of spatial room impulse response extrapolation by direct and residual subspace decomposition," in *AES Int. Conf. on Audio for Virtual and Augmented Reality*, Redmond, WA, 2022.

[17] Rothauser, E., "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio and Electroacoustics*, 17(3), pp. 225–246, 1969.

[18] McKenzie, T., Murphy, D., and Kearney, G., "Assessing the authenticity of the KEMAR mouth simulator as a repeatable speech source," in *143rd AES Conv.*, New York, 2017.

[19] Kabal, P., "TSP speech database," Technical report, McGill University, Database Version, 2002.

[20] McKenzie, T., McCormack, L., and Hold, C., "Dataset of spatial room impulse responses in a variable acoustics room for six degrees-of-freedom rendering and analysis," in *arXiv preprint*, pp. 1–3, 2021, doi:10.48550/arXiv.2111.11882.

[21] Zaunschirm, M., Schörkhuber, C., and Höldrich, R., "Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint," *J. Acoust. Soc. Am.*, 143(6), pp. 3616–3627, 2018, doi: 10.1121/1.5040489.

[22] McKenzie, T., Murphy, D. T., and Kearney, G. C., "Diffuse-field equalisation of binaural Ambisonic rendering," *Appl. Sci.*, 8(10), 2018, doi:10.3390/app8101956.

[23] Bernschütz, B., "A spherical far field HRIR / HRTF compilation of the Neumann KU 100," in *Fortschritte der Akustik – AIA-DAGA 2013*, pp. 592–595, 2013.

[24] "ITU-R BS.1534-2: Method for the subjective assessment of intermediate quality level of audio systems," Technical report, International Telecommunication Union, 2015.

[25] Schoeffler, M., Bartoschek, S., Stöter, F. R., Roess, M., Westphal, S., Edler, B., and Herre, J., "webMUSHRA - A comprehensive framework for web-based listening tests," *J. Open Research Software*, 6(1), 2018, doi:10.5334/jors.187.

[26] Hintze, J. L. and Nelson, R. D., "Violin plots: a box plot-density trace synergism," *Am. Statistician*, 52(2), pp. 181–184, 1998.

[27] Holm, S., "A simple sequentially rejective multiple test procedure," *Scandinavian J. Stat.*, 6(2), pp. 65–70, 1979.