Spatial Perception of Multi-Source Scenarios in Real and Virtual Loudspeaker Arrangements

Stefan Riedel[1], Matthias Frank[1]

[1] *Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, 8010 Graz, Austria*

Email: riedel@iem.at, frank@iem.at

Introduction

In literature, many studies compared the perceptual quality of non-individual vs. individual binaural rendering methods, oftentimes concluding that non-individual head-related transfer functions (HRTFs) of dummy heads deliver a high overall quality, comparable or better than individually-measured HRTFs [1, 2, 3, 4]. However, non-individual HRTFs can lead to increased localization errors, particularly in the vertical dimension [5, 6, 7]. The present study investigates the perception of multi-source scenarios in an ecologically-valid environment including visual cues of a loudspeaker arrangement, cf. Fig. 1. The aim is to compare non-individual dynamic binaural rendering with loudspeaker reproduction. In particular, we investigate the perception of (1) distance and elevation as low-level attributes, and (2) listener envelopment (LEV) and engulfment (LEG) as high-level attributes [8, 9]. While some differences are expected for the low-level attributes, it is unclear whether envelopment or engulfment are affected by non-individual binaural rendering. Open headphones are employed during the experiments, which allow in-situ presentation of real and virtual stimuli [10].

Experiment Design

Spatial Attributes. The attributes tested in the experiment are partly described in the spatial audio quality inventory (SAQI) [11]: distance, vertical localization (elevation), and envelopment. Additionally, 'engulfment' is tested here, which is typically defined as the 'sensation of being covered by sound' [8, 9].

Room and Loudspeaker Setup. The experiments were conducted in a studio room equipped with a 21-channel loudspeaker array (room dimensions of $6.2\,\mathrm{m} \times 4.3\,\mathrm{m} \times 3.4\,\mathrm{m} = 90.6\,\mathrm{m}^3$), cf. Fig. 1. The reverberation time of the room is estimated to $\mathrm{RT}_{30} = 0.204\,\mathrm{s}$ as the octave-band average from 250 Hz to 8 kHz, measured in the center position of the array using an omnidirectional Earthworks measurement microphone. The loudspeakers employed were Genelec 8020 full-range loudspeakers, specified by the manufacturer to have a ±2.5 dB flat frequency response from 60 Hz to 20 kHz.

Dynamic Binaural Rendering. The dynamic binaural rendering used high-resolution, diffuse-field-equalized HRIRs of the KEMAR and KU100 dummy heads (2×2 degree resolution in azimuth and elevation). The direct sound was rendered in six degrees of freedom (6-DoF) by convolution with the nearest-available HRIR, incorporating loudspeaker directivity and $1/r$ distance gains (see [12] for implementation details).

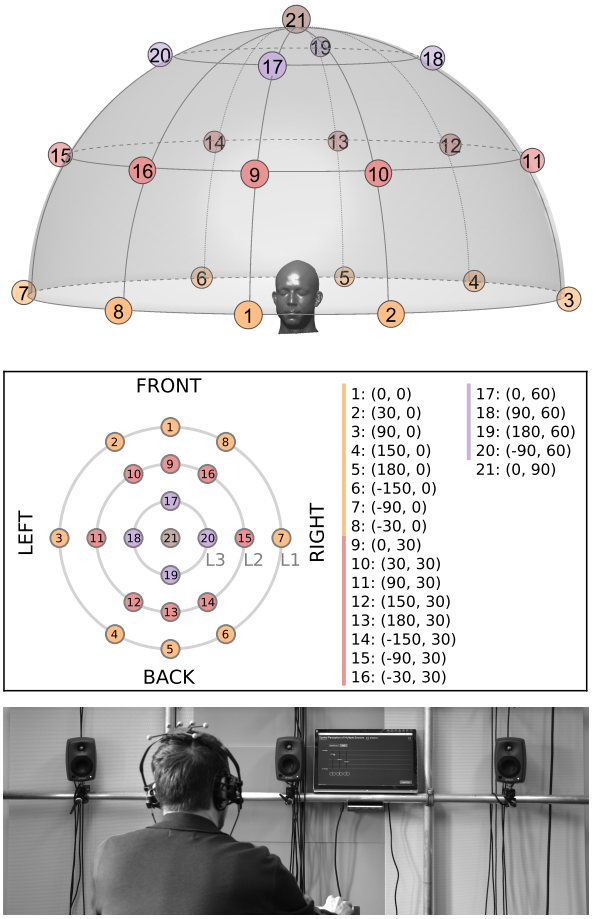

Figure 1: 3-D view of hemispherical loudspeaker arrangement (top) with three elevation layers: 0° (L1), 30° (L2), and 60° (L3), and 2-D projection (middle) with directions listed on the right (azimuth and elevation in degrees, loudspeaker 21 not active). Participant wearing open headphones (bottom).

The room acoustic auralization was implemented in 3-DoF, as participants had to remain in the center of the loudspeaker arrangement throughout the experiment. First-order directional room impulse responses were measured in the center using a tetrahedral Soundfield ST450 microphone, and were subsequently upmixed to 5th-order Ambisonics using the Ambisonic spatial decomposition method (ASDM) [13]. Binaural decoding employed the magnitude-least-squares (MagLS) approach with KU100 HRTFs. Levels of direct sound and room rendering were adjusted with reference to a frontal binaural room impulse response of the KU100 dummy head. Equalization of the open headphones [10] was performed in third-octave bands based on KU100 dummy head measurements.

Table 1: Spatial scenarios in the multi-source conditions. Loudspeaker directions are illustrated in Fig. 1.

Spatial scenario	Active loudspeakers
Frontal	$1, 2, 8$
Wide	$2, 3, 7, 8$
Layer L1	$1, \ldots, 8$
Layer L2	$9, \ldots, 16$
Layer L3	$17, \ldots, 20$
Anchor (for LEV / LEG)	$1, 2, 8$

In the employed binaural rendering, the direct sound had to remain time-aligned with the 3-DoF room acoustic auralization measured for the center position. Therefore, only static delays based on the loudspeaker distances to the center position were applied to the virtual loudspeaker channels, to incorporate the delay-based decorrelation of loudspeaker signals. The phase of the (direct-sound) HRTFs was smoothly transitioned to zero-phase above $f_c = 1.5\,\mathrm{kHz}$, to avoid destructive interference at high frequencies for amplitude-panned sources (cf. time-alignment in HRTF interpolation methods [14]). In the real loudspeaker reproduction, the interference of coherent sources occurs also, but it is affected by the time-of-flight differences among the loudspeakers.

Stimuli. To compare perceived distance, anechoic recordings of a string quartet ensemble [15] were assigned among loudspeakers in the frontal area, cf. Tab. 1 (row 1). Specifically, the first violin was assigned to channel 2, the second violin and viola to channel 1 (center), and the cello to channel 8. In a second scenario, a frontal speech stimulus was rendered using a 5th-order Ambisonic granular synthesis ('GranularEncoder' [9]), which distributed uncorrelated speech syllables uniformly random between $-30°$ and $+30°$ azimuth in the horizontal layer. The Hann-windowed grains of duration $L = 250\,\mathrm{ms}$ were randomly extracted from a female speech sample and spatialized every $\Delta t = 20\,\mathrm{ms}$. To be consistent with the channel-based string quartet condition, channel signals apart from loudspeakers 1, 2, and 8 were set to zero after Ambisonic decoding [16].

To test listener envelopment (LEV), a wide string quartet condition was rendered by assigning the four audio tracks to channels 2, 3, 7, and 8 in the horizontal layer, cf. Tab. 1 (row 2). Furthermore, a uniformly surrounding speech stimulus was rendered using the Ambisonic granular synthesis processing ($L = 250\,\mathrm{ms}$, $\Delta t = 20\,\mathrm{ms}$). All channels apart from layer L1 were set to zero after Ambisonic processing, cf. Tab. 1 (row 3).

Elevation and engulfment (LEG) were tested for different loudspeaker layers. In particular, pink noise grains ($L = 20\,\mathrm{ms}$) or female speech grains ($L = 250\,\mathrm{ms}$) were distributed to the L2 layer (30° elevation) or L3 layer (60° elevation), cf. Tab. 1 (row 4 and 5). To maximize localizability for the pink noise grains, directional overlap was removed by setting the grain duration to $L = 20\,\mathrm{ms}$ at a spatialization interval of $\Delta t = 20\,\mathrm{ms}$.

Table 2: Spatial attributes investigated in the listening experiment. Distance and elevation are low-level attributes, whereas envelopment and engulfment are high-level attributes.

Attribute	Scale
Distance	'In-head' to 'Loudspeaker' (≈ 2 m)
Elevation	$0°$ to $90°$
Envelopment	0 to 100 ('surrounded by sound')
Engulfment	0 to 100 ('covered by sound')

Experiment Protocol. Participants rated the different rendering methods in a multiple-stimulus comparison, which allowed them to switch repeatedly between the three blind conditions (real loudspeakers, KEMAR rendering, KU100 rendering). For envelopment and engulfment, a frontal scenario was added as hidden anchor and reproduced over the real loudspeakers, cf. Tab. 1 (row 6). A summary of the four attributes and the rating scales can be found in Tab. 2.

Subjects. Fourteen participants took part in the experiment ($N = 14$, two females, mean age of 34 years). Ten of the subjects can be considered expert listeners due to participation in multiple spatial audio experiments.

Results & Discussion

The results of the experiment are shown in Figs. 2 to 5. Pairwise Wilcoxon signed-rank tests are conducted to compare the real loudspeaker playback (wearing open headphones) with the two virtual rendering conditions. Horizontal lines printed in black indicate statistically significant differences as of $p < 0.05$ (otherwise printed gray). The effect size $|r|$ is reported for each comparison ($0 \leq |r| \leq 1$, where $|r| > 0.3$ is considered a moderate effect and $|r| > 0.5$ is a strong effect [17]).

Distance. The median distance rating for the loudspeaker playback is higher than for the two virtual rendering variants in the frontal string quartet scenario, but the difference is not statistically significant, cf. Fig. 2 (left). The effect size is small for the comparison between loudspeaker and KEMAR rendering ($|r| = 0.25$) and moderate for the KU100 rendering ($|r| = 0.43$). For the speech stimulus, the differences are less pronounced with small effect sizes $|r| \leq 0.23$, cf. Fig. 2 (right).

Envelopment. Regarding listener envelopment, the wide string quartet scenario reveals significant differences among real and virtual rendering, cf. Fig. 3 (left). Interestingly, it is the virtual conditions that achieve very high median ratings, clearly above the loudspeaker rendering. The occlusion of the open headphones for lateral sources at $\pm 90°$ azimuth can be assumed to be responsible for a degradation in the quality of the real loudspeaker stimulus (see [18, Fig. 2g]). In contrast, for the diffuse granular speech, the median rating of the real loudspeaker rendering is higher than for the virtual rendering conditions, cf. Fig. 3 (right). The effect sizes are moderate with $|r| = 0.42$ for the KEMAR rendering and $|r| = 0.31$ for the KU100 rendering, but differences are not statistically significant ($p > 0.05$). Overall, the virtual rendering meth-

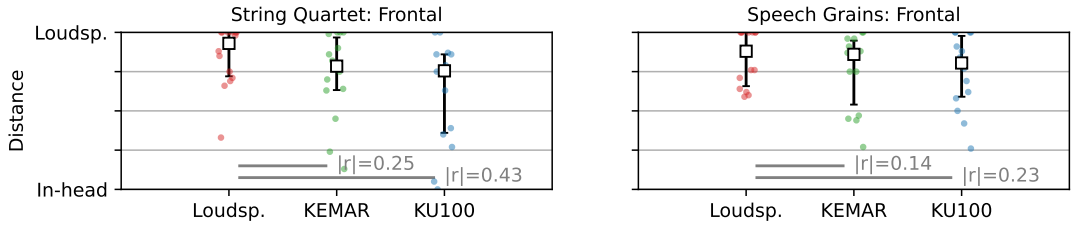

Figure 2: Median and interquartile range for perceived distance ($N = 14$ participants). Horizontal lines indicate pairwise Wilcoxon signed-rank tests, printed in black for $p < 0.05$ and gray otherwise. Additionally, the effect size $0 \leq |r| \leq 1$ is reported.

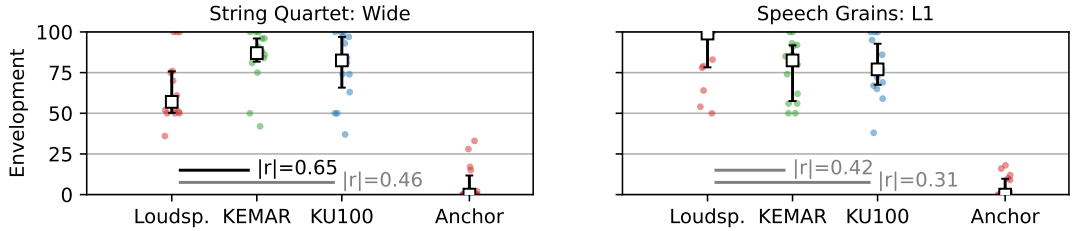

Figure 3: Median and interquartile range for envelopment ($N = 14$ participants). Horizontal lines indicate pairwise Wilcoxon signed-rank tests, printed in black for $p < 0.05$ and gray otherwise. Additionally, the effect size $0 \leq |r| \leq 1$ is reported.

ods achieved high ratings in envelopment with median ratings above 75 out of 100 scale points.

Elevation. The results for perceived elevation are shown in Fig. 4. For the frontal multi-source scenarios (left), the virtual rendering methods based on the KEMAR and KU100 HRTFs reveal a significant upward bias ($p < 0.05$ with $|r| > 0.8$). In particular, the elevation ratings seem to cluster at 30°, which corresponds to the visible L2 elevation layer, cf. Fig. 4 (left) and Fig. 1. The upward bias is clear for the pink noise grains as well as for the speech grains, which confirms and extends recent findings on vertical localization with dummy head HRTFs [6, 19]. The median ratings for the L2 scenario seem accurate for all reproduction methods, with only small effects of downward bias for the virtual rendering conditions, cf. Fig. 4 (middle). For the L3 layer at 60° elevation, the virtual conditions demonstrate a downward bias, cf. Fig. 4 (right).

Engulfment. The results for engulfment are shown in Fig. 5. While a reduction in engulfment can be seen for the virtual rendering compared to the real loudspeaker rendering for the pink noise grains, no clear trend is observable for the speech grains. Clearly, both the virtual and real conditions were perceived as 'covering from above' compared to the frontal anchor condition (channels 1, 2, and 8), cf. Tab. 1.

Conclusion

The presented experiment compared the perception of multi-source scenarios for real vs. virtualized loudspeaker rendering. Results indicate that perceived distance and elevation were impaired in the frontal area for the non-individual dynamic binaural rendering using KEMAR or KU100 dummy head HRTFs. The high-level attributes of envelopment and engulfment could be reproduced with high quality in the virtual rendering, and were either not impaired at all or only slightly impaired, depending on the stimulus condition. The results underline previous studies: overall quality is high with non-individual dummy head HRTFs, but accurate spatial mapping benefits from individual HRTF cues.

Acknowledgment

The authors gratefully received funding from the Austrian Science Fund (FWF) under grant number P 35254-N.

References

[1] C. Armstrong et al., "A perceptual evaluation of individual and non-individual hrtfs: A case study of the sadie ii database," *Appl. Sci.*, vol. 8, no. 11, 2018.

[2] O. S. Rummukainen et al., "Head-related transfer functions for dynamic listeners in virtual reality," *Appl. Sci.*, vol. 11, no. 14, 2021.

[3] M. Blau et al., "Toward realistic binaural auralizations," *Acta Acust.*, vol. 5, 2021.

[4] F. Schultz et al., "Hrtf individualised mag-ls and compass ambisonics-to-binaural rendering: A perceptual evaluation of overall quality," in *Proc. of the 10th Conv. of the EAA*, Turin, Italy, 2023.

[5] E. M. Wenzel et al., "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 111–123, 1993.

[6] S. Riedel et al., "Localization of real and virtual sound sources in a real room: effect of auditory and visual cues," *Submitted to J. Acoust. Soc. Am.*, 2024.

[7] Z. Ben-Hur et al., "Localization of virtual sounds in dynamic listening using sparse hrtfs," in *Proc. of the AES Int. Conf. on Audio for Virtual and Augmented Reality*, Online, 2020.

[8] R. Sazdov et al., "Perceptual investigation into envelopment, spatial clarity, and engulfment in reproduced multi-channel audio," in *Proc. of the 31st Int. AES*

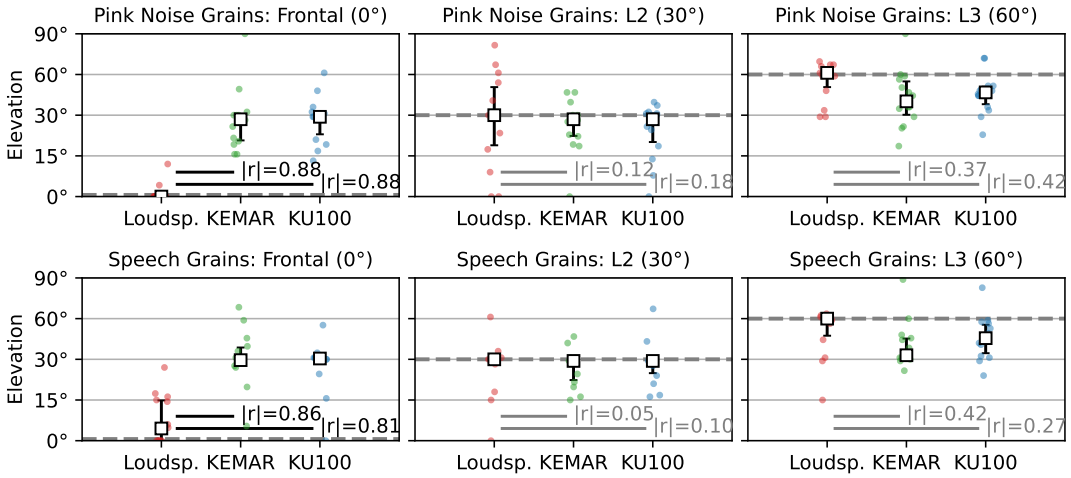

Figure 4: Median and interquartile range for perceived elevation ($N = 14$ participants). Horizontal lines indicate pairwise Wilcoxon signed-rank tests, printed in black for $p < 0.05$ and gray otherwise. Additionally, the effect size $0 \leq |r| \leq 1$ is reported.

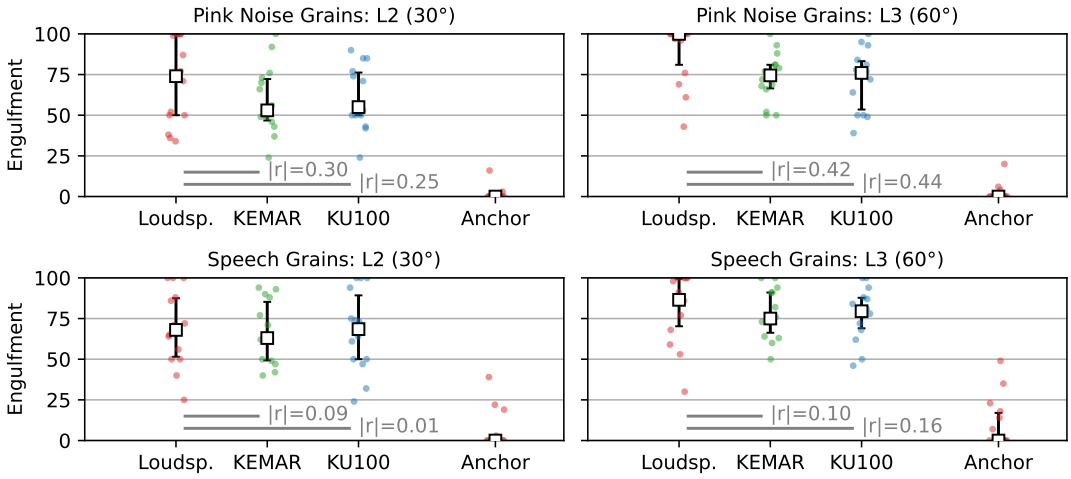

Figure 5: Median and interquartile range for engulfment ($N = 14$ participants). Horizontal lines indicate pairwise Wilcoxon signed-rank tests, printed in black for $p < 0.05$ and gray otherwise. Additionally, the effect size $0 \leq |r| \leq 1$ is reported.

Conf.: New Directions in High Resolution Audio, London, UK, 2007.

[9] S. Riedel et al., "The effect of temporal and directional density on listener envelopment," *J. Audio Eng. Soc.*, vol. 71, no. 7/8, pp. 455–467, 2023.

[10] A. Mülleder et al., "Ultralight circumaural open headphones," in *Proc. of the 154th AES Convention*, Helsinki, Finland, May 2023. [Online]. Available: www.aes.org/e-lib/browse.cfm?elib=22075

[11] A. Lindau et al., "A spatial audio quality inventory (saqi)," *Acta Acust. united Acust.*, vol. 100, no. 5, pp. 984–994, 2014.

[12] S. Riedel, "Hrir convolver vst plug-in," 2023. [Online]. Available: git.iem.at/audioplugins/IEMPluginSuite/-/tree/HRIRConvolver

[13] M. Zaunschirm et al., "Binaural rendering with measured room responses: First-order ambisonic microphone vs. dummy head," *Appl. Sci.*, vol. 10, no. 5, p. 1631, 2020.

[14] J. M. Arend et al., "Assessing spherical harmonics interpolation of time-aligned head-related transfer functions," *J. Audio Eng. Soc.*, vol. 69, no. 1/2, pp. 104–117, 2021.

[15] O. C. Gomes et al., "Anechoic multi-channel recordings of individual string quartet musicians," in *Proc. of I3DA*. IEEE, 2021, pp. 1–7.

[16] F. Zotter and M. Frank, "All-round ambisonic panning and decoding," *J. Audio Eng. Soc.*, vol. 60, no. 10, pp. 801–820, 2012.

[17] C. Fritz et al., "Effect size estimates: current use, calculations, and interpretation." *J. Exp. Psychol. Gen.*, vol. 141, no. 1, p. 2, 2012.

[18] A. Mülleder et al., "Do-it-yourself headphones and development platform for augmented-reality audio," in *Proc. of AES 2023 Int. Conf. on Spatial and Immersive Audio*, Huddersfield, UK, 2023.

[19] M. Frank and S. Riedel, "Simulation study on the effect of (non-)individual hrtfs and ambisonics on median plane localization," in *Proc. of the DAGA*, vol. 50, Hanover, Germany, 2024.