Fidelity of focal sound source reproduction by higher (9th) order ambisonics and perceptual effects of reproduction errors

Nima Zargarnezhad[1,2, a], Bruno Mesquita[2], Ewan A Macpherson[3,4], Ingrid Johnsrude[1,2,3,4]

[1] *Department of Psychology, Western University, London, ON, Canada*

[2] *Western Center for Brain and Mind, Western University, London, ON, Canada*

[3] *School of Communication Sciences and Disorders, Western University, London, ON, Canada*

[4] *National Center for Audiology, Western University, London, ON, Canada*

Higher-order ambisonic rendering is an increasingly common acoustic field reproduction technique that enables presentation of virtual sounds at nearly any location in 3D space, relatively unconstrained by the veridical locations of loudspeakers. We evaluated whether 3D sound reproduction through a 9th order ambisonic system was sufficiently accurate to probe the limits of human spatial perception. In Experiment 1, we estimated minimum audible angles for human listeners at a variety of reference points on the horizontal plane. Our estimated values are similar to absolute thresholds obtained using single-channel free-field presentation for locations within ±51° on the horizontal plane, including values approaching 1° on the midline[1], regardless of speaker density. This demonstrates the adequacy of the AudioDome for studies of human auditory spatial perception, at least for displacement in the horizontal plane. In Experiment 2 we estimated monaural and binaural localization cues by presenting linear chirp sweeps (0-22050 Hz) at the horizontal reference points tested in Experiment 1 and recorded ear canal signals through a head and torso simulator. Although localization cues for low-frequency components were well preserved,

[a] Email: nzargarn@uwo.ca

they were distorted for components above 4000 Hz. In Experiment 3 we provide evidence that these high-frequency distortions are processed as cues to elevation.

Keywords: Higher-Order Ambisonics, Soundscape Reproduction, Sound Localization, Minimal Audible Angle, Spatial Hearing

I. INTRODUCTION

Human auditory perceptual experience is essentially spatial in nature, despite the fact that, unlike for vision and touch, the sensory epithelium is not spatiotopically organized. Whereas our skin and retinas are equipped with receptors that, by their location in the sensory array, encode information about the spatial position of events in the world. Our cochlea is organized by frequency, and appears to compute location from a variety of different temporal and spectral sound features often requiring integration across the two ears. Sound location relative to the head is primarily estimated through binaural cues, such as Interaural Time Difference (ITD) and Interaural Level Difference (ILD) which index the delay of sound arrival time and level between ears[2,3]. While spectral detail in reasonably broadband sounds provides cues to elevation, ITDs are more useful for lower frequencies (below 1000 Hz), whereas ILDs are more salient at higher frequencies (above 1500 Hz)[2,4,5]. In addition to these cues, which inform the listener about the lateral location of a sound source relative to the midline, monoaural spectral cues contribute to estimating the elevation and front/back location of the source. These cues are the result of sound filtering by the listener's head, upper body, and by the pinnae of the ears. This filtering, referred to as the Head-Related Transfer Function (HRTF), is an individual response function that characterizes the sound received by each ear. A pair of HRTFs for a given individual can be used to filter a sound such that it is perceived by that person as coming from a particular location in 3D space. Individual HRTFs enable the recreation of a spatial hearing experience in experimentally controlled settings, over headphones[2,6–8]. In practice, because estimating individual HRTFs is a time-consuming, expensive, complex (requiring specialized equipment) and somewhat unpleasant process (usually involving the insertion of a microphone near the eardrum), generic HRTFs are typically used[3,8]. These simulations, while delivering a 'spatialized' percept, generally do not deliver as compelling a spatial percept as free-field listening.

High-density 3-dimensional loudspeaker (LSP) arrays are expensive, but deliver a compelling, free-field, spatial percept, derived from the participant's HRTFs, and afford opportunities to study naturalistic behaviors (such as audio-visually guided grasping). In the basic implementation, each sound source is presented from exactly one LSP. This method, Single-Channel (SC) panning, limits the presented sound source locations to the locations of actual LSPs - which are typically many centimeters apart. This is obviously inadequate for the study of the limits of human spatial perception. Another method, called Vector Based Amplitude Panning (VBAP), overcomes the limited resolution of SC panning by distributing a virtual sound source's energy among the surrounding LSPs and enables the simulation of sound sources between LSP locations. VBAP first determines the minimum number of nearby LSPs for any given location then distributes the sound gain among those LSPs such that their vector sum results in a sound image with the appropriate gain at the intended location[9]. This algorithm essentially utilizes one LSP when the intended sound source coincides with an LSP location (identical to SC presentation in this situation); two LSPs when it aligns on the line between two LSPs; and three LSPs if the sound location neither coincides with an LSP nor aligns with the line connecting two LSPs. The energy spread (spatial blurring) of a sound source using VBAP presentation depends on its distance from nearby LSPs and LSP geometry, and so sound source images are nonhomogeneous in their apparent diffusion[10].

Ambisonic panning is another reproduction method that utilizes a 3D LSP array to reproduce 3D soundscapes. For ambisonic panning, the soundscape (comprising one sound, or many) is decomposed into a series of spherical harmonics that describe the sound pressure level, velocity, and its higher order derivatives in different directions within the 3-dimensional acoustic environment. Although utilizing higher order derivatives of the sound wave velocity enables the encoding of more focal sound sources, decoding them in the virtual acoustic environment requires more presentation channels (LSPs); therefore, the spatial resolution of ambisonic reproduction is limited by the number

of LSPs in the array. Theoretically, ambisonic panning has higher spatial resolution than SC and VBAP presentation methods with an equalized energy spread using all loudspeakers for all sound source positions[11]. With lower order ambisonic panning systems (orders 1 to 4), rendered sound sources are perceived as "blurry" or diffuse, and resolution is poor. As the order of the system increases, rendered sound sources are, in theory, less blurry, such that a normal listener might perceive them as more crisply focal and thus be better able to resolve closely adjacent sources[12].

Generally, all sound reproduction methods are an approximation of real-world acoustic characteristics and reproduction is not perfect. The perceptual effects of reproduction errors and the extent to which these errors are acceptable depends on the intended use of the system. Presenting sound sources with SC will minimize errors in localization, but it limits the spatial resolution of soundscape reproduction. VBAP has greater spatial resolution than SC, but the energy of sound sources is spread unequally throughout the 3D space, such that the blurriness of sound sources depends on their location relative to LSPs. Although rendering through ambisonics improves on SC and VBAP, even higher order systems may not perfectly reproduce sound locations: errors in timbral and reconstructed ITD and ILD information have been noted in simulations[13]. Additionally, soundscape reconstruction with ambisonics is valid within the area in which the superposition of spherical harmonics is reliable; a minimum ambisonics order is required for the system to reproduce planewaves below a particular wavelength reliably, within a sphere of a particular size located at the center of the array.[7,14] The possibility of reproduction errors affecting perception may preclude the use of this technology for spatially precise reproduction of acoustic environments..

The goal of the present study is to validate the utility of a custom-built LSP array for conducting studies of human perception of auditory space. This array at Western University (called the "AudioDome", Sonible, GmbH, Austria) consists of 4 dual-channel subwoofers and 91 LSPs

geodesically arranged in a sphere with a radius of approximately 1.65 m, located in a sound attenuating chamber (FIG 1.A). The AudioDome is capable of single-channel (SC), VBAP, and 9th-order ambisonic panning, in order to reproduce 3D acoustic environments. We aimed to characterize the spatial fidelity and perceptual outcomes of the AudioDome's ambisonic panning to validate its use, and identify any limitations, before future psychoacoustic and soundscape studies are conducted using the setup, and this work is also valuable for others using similar Higher-Order Ambisonic (HOA) arrays for human perceptual research.

We first measured horizontal spatial acuity by human listeners, in order to evaluate the perceptual effects of ambisonic spatial blurring in the horizontal plane. Traditional SC presentation has determined that minimal audible angles (MAA) are ~1° on the midline, and then acuity monotonically decreases to ~2° at 45° azimuth, and to more than ~8° at azimuth locations above 75° off the midline.[1,2] In Experiment 1, we measured listeners' horizontal MAAs. We also investigated whether obtained MAA depended on the distance of the rendered source from the nearest LSPs: higher MAAs in virtual positions further away from LSPs would indicate that the focality of rendered sources was not the same everywhere on the surface of the sphere defined by the LSP array.

Subsequently, using a Head-and-Torso Simulator (HATS), we compared the spectral composition of rendered sound, and reconstructed ITD and ILD cues delivered to the center of the array (Experiment 2), for SC, VBAP and ambisonic panning.

Finally, based on the findings in the first two experiments, we measured the effects of spectral distortions introduced by ambisonics on the perception of elevation by human listeners (Experiment 3).

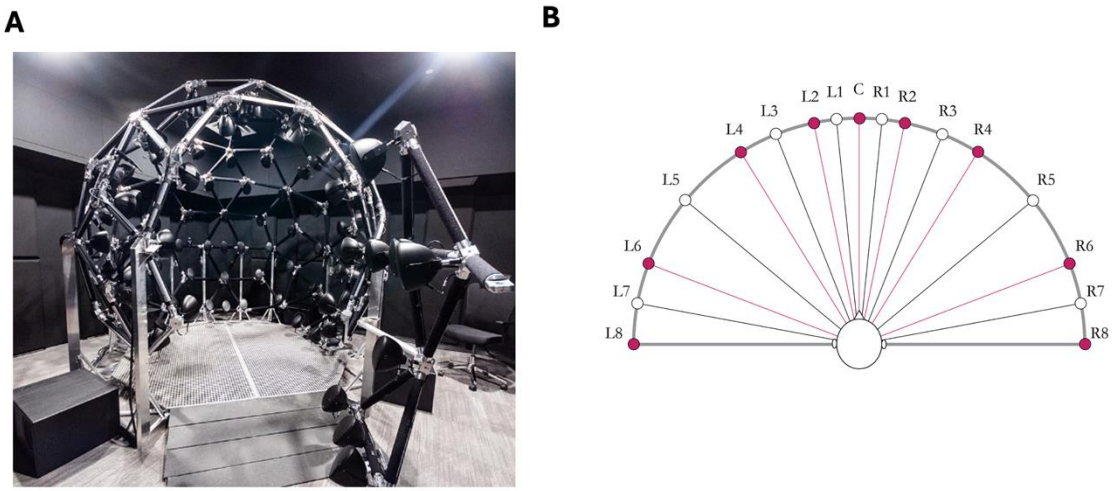


FIG. 1. (color online) A. The AudioDome (Sonible, GmbH, Austria): the array consists of 91 LSPs and 4 subwoofers geodesically arranged in a spherical shape with a radius of approximately 1.65 m in a sound attenuating chamber. This system is equipped with SC, VBAP, and 9th order ambisonic panning technologies. The AudioDome in a sound attenuating chamber. B. A schematic of the reference points used in Experiments 1-3. Locations of LSPs (filled circles) and locations between LSPs (open circles) on the horizontal plane with their labels and the participant's head at the center.

II. EXPERIMENT 1

The first experiment examined whether the spatial blurring of sound sources reproduced with 9th-order ambisonics was above or below the threshold of human acuity (MAA) studied with SC reproduction[1,2,15]. Furthermore, we examined whether spatial acuity was worse (indicative of more blurred sources) when virtual sources were located at a greater distance from the nearest LSPs.

We estimated Minimum Audible Angles (MAAs)[1,2,15] on the frontal half of the horizontal plane at several locations that coincided with actual LSPs or at the midpoints of adjacent pairs (FIG 1.B). The highest spatial acuity (lowest MAAs) was expected at the front of the array (0° azimuth)[1,2] and should approach 1° if ambisonic panning renders point sources at a spatial resolution at or below the limits of human acuity. However, the LSP density is also greatest at 0° azimuth. Accordingly, we

112 tested another condition in which participants were rotated to face another region of the array with
113 the lowest LSP density, also on the horizontal plane. This allowed us to determine whether distance
114 from the nearest LSPs changes the focality of sources. If we observe MAAs of ~ 1° in both
115 conditions, we can conclude that the array is able to render point sources at a spatial resolution at or
116 below the limits of human acuity, everywhere on the surface of the sphere defined by the geodesic
117 array.

118 **A. Materials and Methods**

119 Participants were seven (4 female) young (aged 18-25 years), healthy, normally hearing
120 (thresholds measured audiometrically at both ears at octaves of 125 Hz up to 8 kHz; participants
121 were excluded if their threshold exceeded 25 dB HL at any tested frequency in either ear), right-
122 handed, naive listeners with no history of neurological disorders. Participants sat in the middle of the
123 AudioDome with their head positioned at the center of the dome. The chair was set to face the
124 origin of the array's coordinate system (0° azimuth and elevation angles) so that the ears aligned with
125 the ±90° azimuth line, and the ear level was aligned with the 0° elevation level by adjusting the chair
126 height.

127 Horizontal MAAs were estimated at 17 locations (reference points) on the horizontal plane; nine
128 of these coincided with LSPs within the ±90° azimuth range and eight were at midpoints between
129 adjacent LSPs (FIG 1.B and TABLE I). Participants completed a location discrimination task in
130 which they reported whether the second noise burst of a pair was in a location clockwise, or
131 counterclockwise, from the first. In each trial, one sound was at the location of a reference point,
132 and one at the location of a test point. In each of 17 experimental blocks, one reference point was
133 tested against 13 test points (six on each side, and one co-located) 20 times each. These data were
134 then used to calculate a psychometric function relating the probability of a perceived direction

135 response to the distance of test points from the reference. Test points were presented across larger
136 ranges for the four most lateral reference points given that acuity is poorer in this region[1]; test points
137 ranged across ±12° from the reference for L8 and R8, ±9° for L7 and R7, and ±6° for all other
138 reference points (to maintain the number of test points at 13, the increment size was adjusted
139 accordingly, all ranges are inclusive). The MAA at each reference point was estimated by fitting a
140 sigmoid to each individual's psychometric data, and then reading off the 75% threshold value, which
141 was taken as the just-noticeable difference (JND[16]) . Equation 1 describes the model $\psi_{RP}(\varphi)$ used to
142 describe the probability ,ψ, that the test sound, rendered at probe angle φ, was perceived to be
143 counterclockwise relative to the sound rendered at the reference point, RP. Parameters α_{RP} and β_{RP},
144 representing the center and steepness of the sigmoid function, were estimated using a non-linear
145 least-squares method[17,18].

146 $$\psi_{RP}(\varphi) = \frac{1}{1+e^{\frac{-(\varphi-\alpha_{RP})}{\beta_{RP}}}} \quad (1)$$

147 The two sounds (reference and test) were broad-band white-noise bursts (range 0-22050 Hz at
148 44100 Hz sampling rate), 450 ms in duration, separated by a 200 ms silent gap[15]. Participants could
149 respond (and change their response) up to 1250 ms after the second burst ended. The noise bursts
150 were generated using the 'wgn' function in MATLAB 2022b software[19] with the signal power set to -
151 25 dBW with 22.5 ms long cosine^2-shaped attack and decay ramps. The audio files were presented at
152 73.1 dB SPL to the participants in the AudioDomeAudioDome, which had a background noise level
153 of 25.2 dB SPL at its quietest (SNR = 47.9 dB). Eighty different noise bursts were generated and
154 pseudo-randomly assigned to different trials in the experiment.

155 The experiment was held in the AudioDome (Sonible GmbH, Vienna, Austria), the loudpseaker
156 array that consists of 91 KEF E 301 loudspeakers with 4.25" coax drivers in egg-shaped drivers and
157 approximately linear frequency responses, and 4 subwoofers with 2 independent drivers for the

158 lower portion of the spectrum (<150 Hz), geodesically arranged in a spherical shape with a radius of
159 approximately 1.65 m. This system is equipped with SC, VBAP, and 9th order ambisonic panning
160 technologies.

161 The experiment was held in darkness to rule out the influence of vision on sound localization[20];
162 all the surfaces on the chamber walls were covered with black sound absorbing foam and black
163 Velcro was used to cover the AudioDome skeleton and LSP joints in order to minimize potential
164 acoustic and optical reflections. The only light source was the red LED used for eye fixation, placed
165 at 0° azimuth and elevation. This was illuminated for the duration of each trial to inform the
166 participant of the trial progression (illuminated at the start of each trial and extinguished at the end).
167 The participants were instructed not to move, to keep their eyes open, and to fixate on the red light.
168 An 8-mm magnetic sensor was attached behind the participants' heads to track their motion. (Those
169 head tracking data were not used). Participants used two buttons on a small wireless keypad to
170 indicate the direction in which they perceived the sound to 'jump' from the first burst to the second
171 (clockwise/counter clockwise) on each trial. The order of test and reference points was
172 counterbalanced, such that for each reference point, the first burst of each trial was located at the
173 test point on 10 trials, and at the reference point on the other 10.

174 Reference point C, at 0° azimuth and elevation and coincided with an LSP, is also surrounded by
175 several other LSPs; four within ±11.64° azimuth and ±20.91° elevation. In contrast, reference points
176 L5 and R5 are themselves virtual and in rather sparse LSP regions: the nearest LSPs are located at
177 ±18.69° azimuth and±11.64° elevation (FIG 3.A). The higher concentration of LSPs at the front
178 (reference point C) may result in more focal sound source reproduction at this location, enabling
179 more accurate discrimination and lower MAAs. If we only test participants facing towards reference
180 point C, this density difference is confounded with spatial acuity, which is best at the front and
181 decreases at lateral positions[1]. To test the effect of density without the confound, we tested two

182 further conditions (blocks), identical to the others, in which we rotated the chair so that participants
183 faced L5 and R5. This enabled us to compare frontal MAAs for the reference points in the regions
184 of highest (C) and lowest (L5 and R5) LSP concentrations. In these 'rotated' blocks, the participants
185 completed the task for a reference point in front of them (L5 or R5). The red LED fixation point
186 was set at L5 and R5 for these blocks.

187 The 19 experiment blocks (17 facing reference point C, one facing L5, and one facing R5) were
188 split into two sessions over two days. Session One consisted of consent, audiometry, a demographic
189 questionnaire, head measurements (not presented or analyzed here), followed by training, a practice
190 block, and eight blocks testing four pairs of symmetric reference points. In Session Two,
191 participants completed another practice block, then the remaining four pairs of reference points, the
192 two rotated conditions and reference point C. At the end, the participants were asked about their
193 attentiveness during the task, their understanding of the task, and any strategies they used to
194 respond, then they were debriefed. The reference point pairs were randomly assigned to the two
195 sessions at the beginning of the first session. The order of reference point blocks within a session
196 was randomized. Also, the order in which the three frontal MAA blocks at C, L5, and R5 were
197 tested was pseudo-randomly assigned to participants (each of the six possible permutations was
198 assigned to a participant). Within each block, the order of the trials was fully randomized.

199 At the beginning of each session, participants completed a practice block which consisted of 20
200 suprathreshold trials with pairs of noise bursts presented at different locations drawn from the
201 following locations: C, L2, L4, L6, and L8 (the left LSP locations). The leftmost noise burst of the
202 pair was presented first on half of the practice trials (i.e., clockwise response would be correct). To
203 ensure participants properly learned the task, they only proceeded to the main experiment block
204 after achieving 100% accuracy on the practice block for pairs separated by relatively large distances
205 (i.e., C vs. L6, C vs. L8, and L2 vs. L8). Feedback on accuracy was provided during practice blocks;

206 in the main experiment participants were informed about missed trials or too slow reaction times at
207 the end of each block. One participant who marginally passed the audiometry results was not able to
208 achieve a high accuracy after three tries in the first session's practice block and was not tested
209 further. We; we collected and analyzed data from the remaining six participants.

210 **B. Results**

211 Sigmoid functions were adequate fits to data reference points C, L1, R1, L2, R2, L3, R3, R4, L5,
212 and L5 for all 6 participants. At other reference points, one or more participants performed at or
213 below chance level (1 participant at reference points L4, R6, and R7, 2 at L6 and L7, 3 at R8, and 4
214 at L8). As it is impossible to estimate the 75% threshold for a flat line, those data were removed
215 from the analysis. MAA values are illustrated in FIG 2 and summarized in TABLE I. The average
216 MAAs across participants are in the range of 0.95° to 2.38° at reference points within about 32° off
217 the midline (C, L1-L4, and R1-R4; see TABLE I). At reference points that are 50° to 90° from the
218 midline (L5-8; R5-8) these values are more divergent and asymmetrical between left and right. On
219 the left, average MAAs were 4.46° for both L5 and L6, and for the most extreme lateral locations,
220 L7 and L8, they were 5.84° and 21.78°, respectively. On the right, average MAAs were 2.81° at R5,
221 8.48° at R6, 7.94° at R7, and 6.24° at R8. The values at L8 and R8 were based on data from two (L8)
222 or three (R8) participants, and one of the participants included for L8 was an outlier with a threshold
223 of 36.24 degrees. Therefore, these lateral data are probably not reliable.

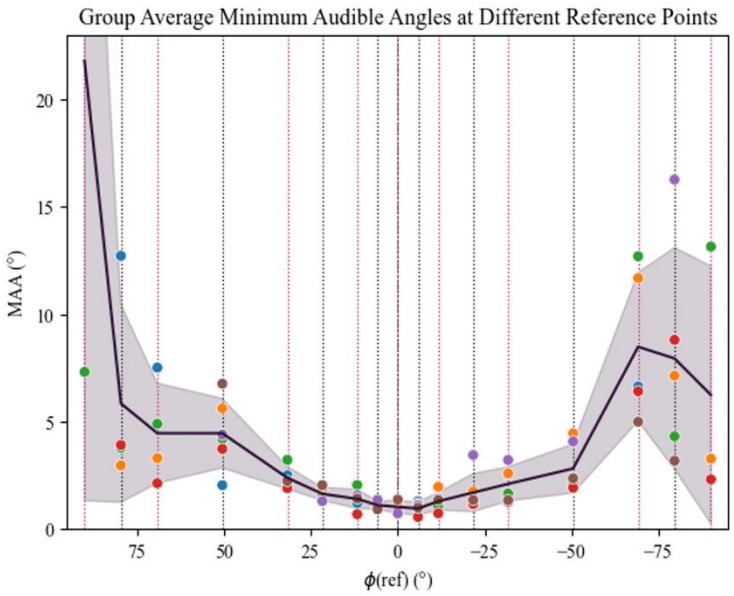


FIG. 2. (color online) The horizontal MAAs for sound sources reproduced with 9th-order ambisonics. The line shows average values across participants; individual 75% thresholds are shown with dots. The shaded region illustrates one standard deviation. (One individual's MAA at -90° (L8) was 36.24°; this value contributed to the average but is not shown in the figure.)

TABLE I. Reference point azimuth coordinate (φ), estimated average MAAs with standard errors (SE), and the number of participants remained in MAA estimation after excluding flat psychometric curves (N). Approximation of previously measured MAAs with pure tones in free-field are read from Fig. 6 of the Mills manuscript[1] and put in the MAA_{SC} column for comparison.

Reference Point		φ_{RP} (°)		MAA ± *SE* (°)		MAA_{SC} (°)	N	
C		0		1.03 ± *0.12*		~1	6	
L1	R1	5.82	-5.82	1.11 ± *0.08*	0.95 ± *0.12*	~1	6	6
L2	R2	11.64	-11.64	1.40 ± *0.18*	1.29 ± *0.17*	~1.2	6	6
L3	R3	21.68	-21.68	1.64 ± *0.12*	1.69 ± *0.36*	~1.5	6	6
L4	R4	31.72	-31.72	2.38 ± *0.23*	2.10 ± *0.32*	~1.8	5	6

L5	R5	50.41	-50.41	4.46 ± *0.66*	2.81 ± *0.47*	~2	6	6
L6	R6	69.09	-69.09	4.46 ± *1.16*	8.48 ± *1.55*	~3	4	5
L7	R7	79.55	-79.55	5.84 ± *2.30*	7.94 ± *2.31*	~4	4	5
L8	R8	90.00	-90.00	21.78 ± *14.47*	6.24 ± *3.46*	> 10	2	3

In the rotated conditions, when participants were facing reference points L5 and R5, the average MAAs were 1.20°±0.08° (N=6) at L5 and 1.17°±0.12° (N=6) at R5. (FIG 3.B). These values do not differ from the frontal MAA obtained at C (three-condition repeated-measures ANOVA; $F(2,10) = 0.8275$, $p = 0.4650$). This indicates that similar MAAs are obtained for frontal positions, regardless of the density of the LSP array at that location.

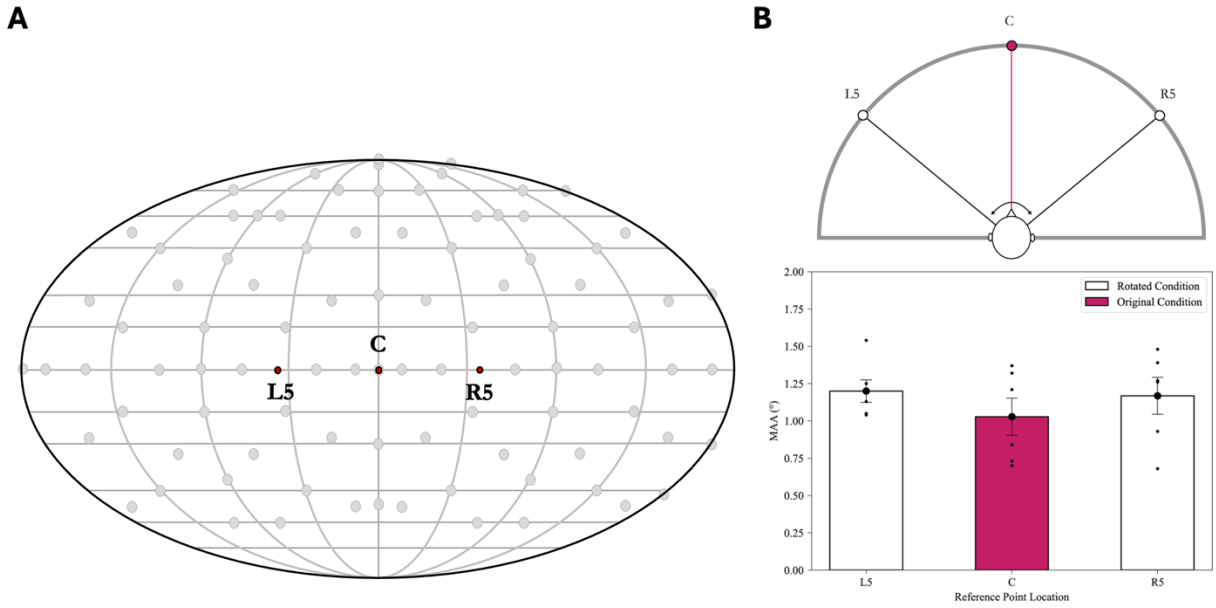


FIG. 3. (color online) A. A flattened illustration of the LSP locations (light grey dots) in the array with reference points C, L5, and R5 (small dark dots). Note that C coincides with an LSP and is surrounded by other LSPs nearby (as close as ±11.64° horizontally), while L5 and R5 are ±18.69° azimuth and ±11.64° elevation away from the nearest LSPs. B. The midline MAA estimated at

243 different locations with different LSP densities. Individual data are illustrated with small dots and the
244 error bars show standard error of the mean.

245 In the post-experiment questionnaire, all participants reported perceiving noise bursts presented
246 at lateral locations L6-L8 and R6-R8 as above or below the horizontal plane, and this was
247 particularly true on the right side. Some participants used this illusory or artifactual elevation percept
248 as an indication of location in the horizontal plane. For example, one participant had concluded that
249 in trials at the R8 reference point, the noise burst that is perceived as higher was located clockwise
250 relative to the other one. (This in fact was wrong and resulted in a flipped psychometric function.)
251 The possible cause of the elevation percept is investigated further in the following experiments.

252 **C. Discussion**

253 In this experiment horizontal MAAs were estimated for 17 reference points on the horizontal plane
254 for wide-band noise bursts reproduced with 9th-order ambisonics. The estimated MAAs were
255 consistently low around the midline and increased at more lateral locations. These results align well
256 with previous reports of human spatial acuity tested in free-field settings: MAAs within the range of
257 ±25° azimuth match with the values reported by Mills [1] measured for 500 Hz and 1000 Hz pure
258 tones[1]. At other locations, the estimated MAAs do not differ substantially from the Mills[1] values
259 except for L8 (90° on the left). Also, all frontal MAAs reported here (in both the original and
260 rotated conditions) were smaller than those obtained in free-field testing with high-pass pink noise,
261 as reported by Strybel and Fujimoto[15], possibly because of the greater sound information and cues
262 available with broad-band noise. Our results demonstrate that sound sources reproduced with 9th-
263 order ambisonic panning are sufficiently focal such that two adjacent sources can be resolved at or
264 below the level of normal human spatial acuity at least for broadband sounds on the horizontal
265 plane, within 50.41° off the midline. The MAAs obtained at lateral azimuthal locations (greater than,

50.41° from the midline) diverge from those recorded using free field methods[1]. The divergence of these may be due to the small number of samples, since several participants were excluded at these lateral locations.

Frontal MAAs were tested on the midline, both when the individual was facing reference point C, where the LSP density was highest, and also when the individual was facing L5 and R5 (see FIG 3), regions with the lowest LSP concentration. Obtained thresholds did not differ, implying that the simulation precision is homogeneous across regions with different LSP densities. This observation reassuringly indicates that the energy spread, and the blurring of virtual sound sources, are truly location independent – at least with the 9th-order ambisonic panning used here. This confirms theoretical predictions about the advantage of higher-order ambisonics over VBAP[7].

Finally, we investigated the anecdotal reports of elevation perception, which were unexpected. We first made sure that the simulation codes did not send any commands to reproduce sounds sources at any elevation level other than 0°. Then, three authors (NZ, BM, and IJ) and another person (DQ) sat in the AudioDome and listened to a block of the task at L8. We agreed that the virtual sound sources were sometimes heard from different elevation levels. This illusion disappeared when we rotated to face L8 in the middle of the block. This observation motivated Experiment 2, in which we quantified the localization cues for virtual sound sources rendered with higher-order ambisonics.

III. EXPERIMENT 2

In Experiment 1, participants' performance dropped substantially at lateral sound sources, more than expected based on previous free-field measurements[1]. For example, at ±90° (reference points L8 and R8), data for at least half of the participants were at chance, or below chance, at all test points Participants in Experiment 1 also reported hearing sounds from lateral locations as above or

289 below the midline. These observations suggest that the spectral profile of the sounds was distorted
290 and was providing artifactual cues to elevation. Given that elevation perception relies on high-
291 frequency monoaural spectral information[2,21], we suspected that the ambisonics algorithm or
292 implementation (erroneously) resulted in the creation of notch filtering in the higher part of the
293 spectrum.

294 To investigate whether elevation perception at lateral locations was perhaps a result of artifactual
295 localization cues, we first evaluated spectral and ITD and ILD information for ambisonic sound
296 sources as recorded through a Head and Torso Simulator (HATS) placed in the same position and
297 facing the same direction as the participants in Experiment 1 (FIG. 4). Then we estimated the
298 spectrum, ITD, and ILD values for wideband-noise sources presented from horizontal LSP
299 locations using SC, VBAP, and ambisonic panning. Then, the values were estimated for the low-
300 frequency and the high-frequency portions of the spectrum and were compared across the two
301 bands. As SC only applies the LSP frequency response to the audio files it serves as the "ground
302 truth" for measurements. Hence, to identify errors in the spectral content of reproduced sound
303 localization cues, ITD, and ILD values derived from recordings of sound presented using VBAP
304 and ambisonic panning were compared to those using SC presentation.

305 **A. Materials and Methods**

306 The HATS type 4128C produced by Brüel & Kjær (Denmark) was used in this experiment. This
307 manikin simulates the acoustic properties of an average adult's upper body, head, and ears. The
308 HATS was placed on a tripod at the position of an ideal listener in the AudioDome and faced
309 towards the origin of the system with the ears at 0° elevation level aligned with the ±90° azimuth
310 line (FIG. 4).

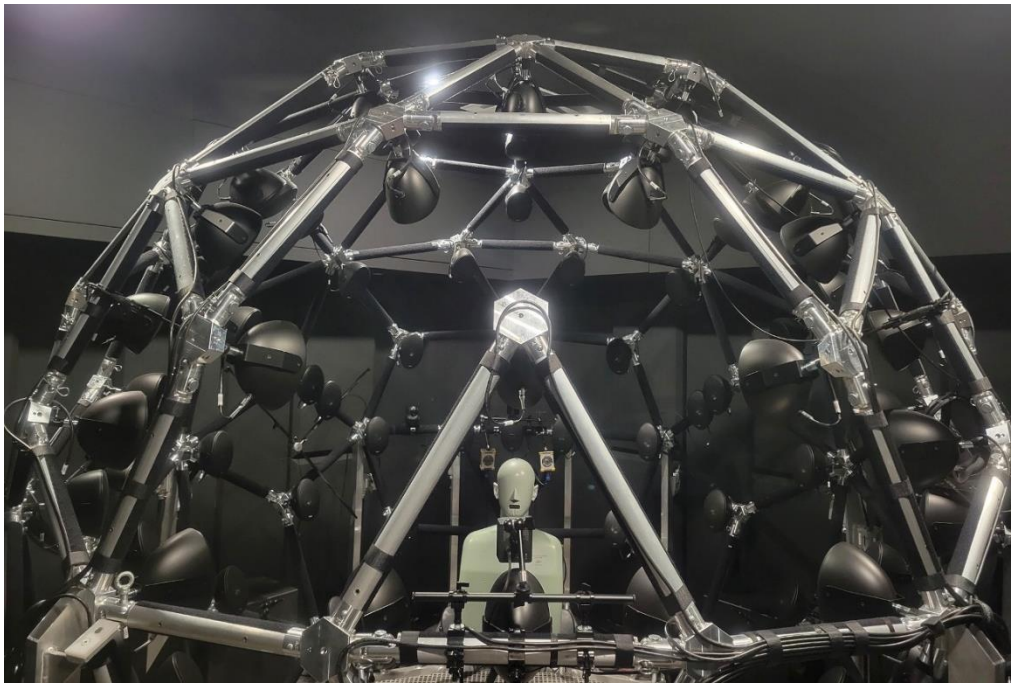


FIG. 4. (color online) The HATS manikin in the measurement position in the AudioDome.

A sinusoidal chirp sweep was presented from target locations while the acoustic content was recorded by two half-inch microphones located in the HATS built-in ears; to estimate binaural and monaural cues as they are heard by an actual (average) human. The chirp signal was a broad-band linear frequency sweep starting from 0 Hz to 22050 Hz sampled at 44100 Hz rate with a duration of 22.05 s (1000 Hz/s) and a constant amplitude of 78.1 dB SPL (Equation 2).

$$x_{chirp}(t) = \sin(1000\pi t^2) \tag{2}$$

The sweep was presented from all the reference points tested in Experiment 1 (TABLE I) with VBAP and ambisonic panning modalities, and SC when possible (when source location coincided with an LSP; i.e., reference points C, L2, R2, L4, R4, L6, R6, L8, and R8). For each method, sweeps were presented 4 times at each location, starting from R8 in counterclockwise order to L8. The sweeps were interleaved with 2 s periods of silence, to ensure that any reverberations from the previous sweep presentation had ceased, and that panning commands and triggers were transmitted properly. Data for all three reproduction methods were recorded in one session in the following

order: SC, VBAP, and ambisonic panning. The HATS position and orientation did not change during the recording.

1. Spectral content estimation

The power spectrum of the chirp sweep, and of each recording, was estimated using the *pspectrum* function in MATLAB[22] with a frequency resolution of 200 Hz and leakage of 0.85. To estimate the HATS transfer function for each panning method, at each target location, in each ear, the spectra of corresponding recordings were divided by the chirp sweep spectrum and averaged across the 4 repetitions. To measure the difference between the original and recorded spectra for high and low frequencies, the spectrum was divided into two bands: a low-frequency band from 0 Hz to 4000 Hz, and a high-frequency band from 4000 Hz to 22050 Hz. The 4000 Hz boundary was chosen based on the equation introduced by Ward and Abhayapala[7,14]; this equation derives the minimum ambisonic order required to accurately reproduce a planewave with a particular wavelength within a sphere of a particular radius at the center of the array. Using this equation, our 9th-order system is capable of accurately reproducing frequencies up to 4000 Hz within a sphere of ~12 cm radius at the center. (Male heads average 56.7 cm in circumference[23], which, if approximated to a circle yields a radius of 9 cm. Thus 12 cm is a conservative estimate that also allows for some error in head position.) Then, the HRTF amplitude spectra of sounds recorded from VBAP and ambisonic panning were compared against SC HRTFs using the gain-difference score ΔG as described in Equation 3[24]. In Eq. 3, $|HRTF_{panning}(f)|$ represents the HRTF magnitude of the 'panning' method at frequency 'f', and 'f_L' and 'f_H' represent the lower and the higher bound of the estimation frequency band, respectively.

$$\Delta G_{panning}(f_L, f_H) = 10log_{10}\frac{\int_{f_L}^{f_H}|HRFT_{panning}(f)|^2 df}{\int_{f_L}^{f_H}|HRFT_{SC}(f)|^2 df} \quad (3)$$

348 ***2. ITD estimation***

349 ITDs are more dominant for frequencies below 1 kHz[2,4,5] and thus any reconstruction error at
350 higher frequencies are not important. However, in the frequency range of 1000-1500 Hz neither
351 ITD nor ILD cues dominate localization[2,4,5], therefore to assess the general quality of ITD
352 reconstruction, the 1000-1500 Hz range with some margin was included in the lower frequency
353 portion of the spectrum for ITD estimations. For ITD measurement, the spectrum was again
354 divided to two portions: but this time the lower frequency band extended from 0 Hz to 1800 Hz,
355 and the higher frequency band from 1800 Hz to 22050 Hz. To estimate the ITD within each range,
356 the portion of the recording that corresponded to each frequency range (based on the chirp signal
357 presentation timing) was taken. Then this portion of the signal was temporally upsampled with a
358 factor of 10 to increase the temporal resolution of the cross-correlation function used in the next
359 stage. Finally, the ITD was estimated as the lag amount that maximized the cross-correlation
360 between the left and right channel recordings (delayed in the right channel relative to the left). ITDs
361 were reported as the average of the 4 values estimated for each location with each reproduction
362 method.

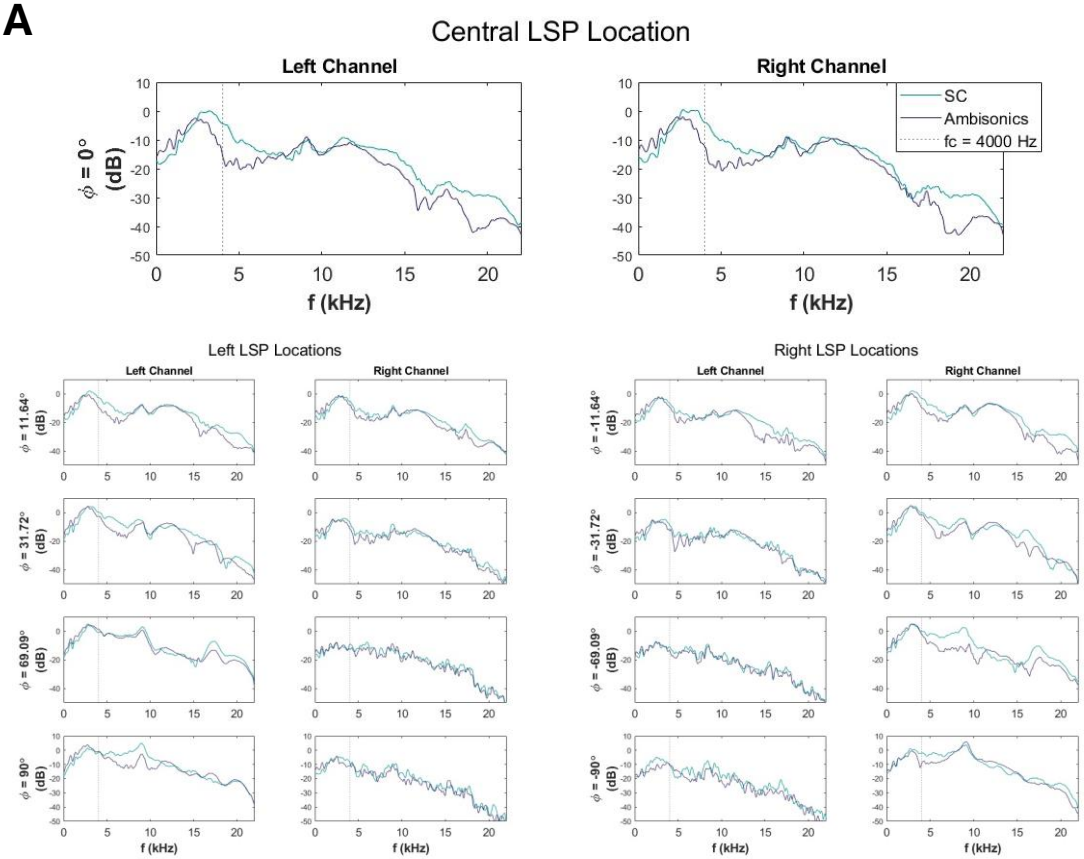
363 ***3. ILD estimation***

364 To estimate ILDs, the Fourier transform of the signal in both ears was estimated with the *fft*
365 function in MATLAB[22]. As in the spectral content analysis, a boundary of 4000 Hz was used to split
366 the spectrum into low-frequency and high-frequency bands. The sum of squared values of the
367 Fourier coefficients of frequencies within each band indicated the spectral power of that band. The
368 ratio of the spectral power in the left channel over the spectral power in the right channel was used
369 to estimate ILDs. ILDs were estimated for each measurement separately and the average across the
370 4 repetitions were used to calculate an ILD value for each reference point and reproduction method.

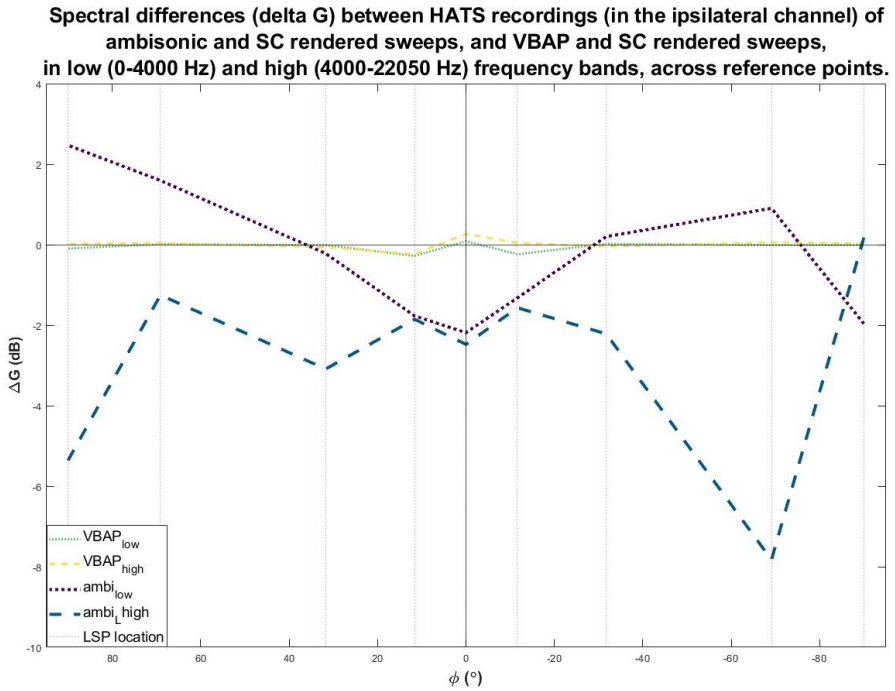
B. Results

The HRTF for the left and right channels of the HATS for sound sources from horizontal LSP locations is shown in FIG 5.A for SC and ambisonic panning methods. This panel, in combination with the gain-difference scores, ΔG, for the two frequency bands illustrated in FIG 5.B, shows that the ΔG is larger at higher frequencies than at lower frequencies, and, overall, ambisonically rendered sounds have lower power than SC sounds. (VBAP is theoretically equivalent of SC panning when the sound sources coincide with an LSP location, therefore VBAP HRTF magnitudes are not included. In panel B the ΔG between VBAP and SC is shown and approximates 0 everywhere, confirming the equivalence of VBAP and SC presentation methods for veridical loudspeaker locations.)

A


Central LSP Location
Left Channel
Right Channel
SC
Ambisonics
fc = 4000 Hz
ϕ = 0° (dB)
f (kHz)
Left LSP Locations
Right LSP Locations
ϕ = 11.64° (dB)
ϕ = 31.72° (dB)
ϕ = 69.09° (dB)
ϕ = 90° (dB)
ϕ = -11.64° (dB)
ϕ = -31.72° (dB)
ϕ = -69.09° (dB)
ϕ = -90° (dB)

B


Spectral differences (delta G) between HATS recordings (in the ipsilateral channel) of ambisonic and SC rendered sweeps, and VBAP and SC rendered sweeps, in low (0-4000 Hz) and high (4000-22050 Hz) frequency bands, across reference points.
ΔG (dB)
ϕ (°)
VBAP low
VBAP high
ambi low
ambi high
LSP location

381

FIG. 5. (color online) A. The HRTF magnitudes (left, and right) for sounds reproduced at LSP locations on the horizontal plane in the AudioDome. Only SC and ambisonic panning methods are shown. The top row shows the spectra of sound sources presented on the midline (reference point C). On the other rows, the left pair of panels shows the spectra of response to presentations at L2, L4, L6 and L8, in the left and right ears. The right pair of panels show the spectra of response to presentations at R2, R4, R6, and R8. B. The gain-difference score (ΔG) which compares the magnitude spectra of SC and other presentation methods is shown for the low frequency (0-4000 Hz) and high frequency (4000-22050 Hz) regions in the channel ipsilateral to the source location (for $\varphi = 0°$ the average of both channels is illustrated). F_c = 4000 Hz was set as the boundary frequency that splits the spectrum. ΔG for VBAP - SC illustrates the noise ceiling of the score - the fact that its values are around 0 dB confirms the equivalence of VBAP and SC for these reference point locations.

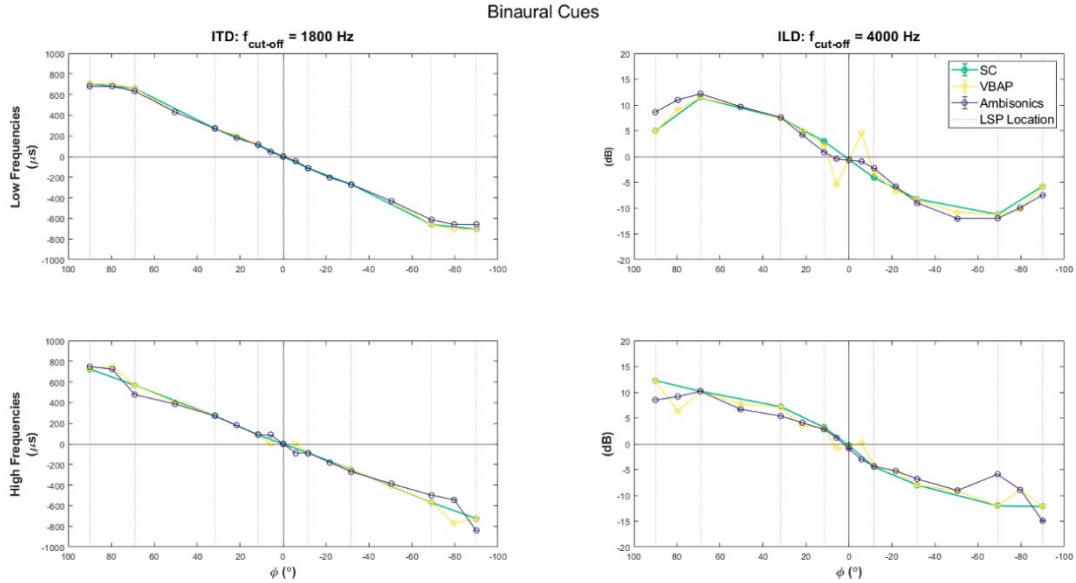


FIG. 6. (color online) The average ITD (left column) and ILD (right column) values in low (top row) and high (bottom row) frequency bands across different reproduction methods for sound

397 sources at reference points across the horizontal plane. The boundary frequencies to split low and
398 high bands were 1800 Hz for ITD and 4000 Hz for ILD panels. Data for locations between LSPs
399 are available only for VBAP and ambisonic panning. Error bars are standard error.

400 For all the reference points tested in Experiment 1, the average ITD and ILD values were
401 estimated for the three sound reproduction methods (only LSP locations for SC) as illustrated in
402 FIG 6. When the sound source was located on the horizontal midline (φ = 0°) ITD values were 0 μs
403 for all reproduction methods, and all repetitions. This confirms that the HATS was positioned
404 properly in the center of the array during the experiment. The average low-frequency (0-1800 Hz)
405 ITDs for SC and VBAP followed a monotonic trend from reference point L8 (–90°) through the
406 midline (C) to R8 (+90°), and ranged between ±703 μs. ITD values were symmetrical between left
407 and right. ITDs measured from VBAP presentations were equal to those of SC ITDs at LSP
408 locations as theoretically expected. For ambisonics, the average low-frequency ITDs ranged from -
409 658 μs to +681 μs with a monotonic trend from reference point L8 (–90°) through the midline (C)
410 to R8 (+90°). ITDs were slightly different between the left and right hemifields (maximum
411 difference ~ 23 μs) at L3/R3, L6/R6, L7/R7, and L8/R8 pairs. The average low-frequency ITDs
412 for VBAP and ambisonics were equal for L7/L8 and R7/R8 pairs. Average high-frequency (1800-
413 22050 Hz) ITDs for SC also followed a monotonic trend within the range of ±726 μs, and had equal
414 magnitudes on the right and left except at the L4/R4 pair where they differed by ~23 μs. Average
415 high-frequency ITDs for VBAP again matched the values of SC at LSP locations and were
416 symmetrical on the left and right, except for the L8/R8 pair which differed by 40 μs. At locations
417 between LSPs,ITDs derived from VBAP data diverged from the monotonic trend and were not
418 always symmetrically equal in magnitude. ITDs measured in the high-frequency region for ambisonic
419 presentation ranged from -840 μs to749 μs, with a non-monotonic trend at the front (L2 to R2); the

magnitude of these ITDs were symmetrically equal up ±51°, and the trend diverged from SC ITD values on the sides (L6-L8 and R6-R8). The standard deviation (across four recordings per reproduction method, per reference point) of all measured ITD values was 0 μs, except for the high frequency band VBAP ITD at $\varphi = -90°$ which had a standard deviation of 11.34 μs.

Average low-frequency (0-4000 Hz) ILDs for SC were within the range of ±11.5 dB with a monotonic trend within ±69.09° azimuth angle and a lower magnitude of ±5.8 dB at ±90°. The average low-frequency ILD was -0.55 dB when the source was at the center (reference point C). (Note that ITD was 0 μs at this point indicating that HATS was probably not poorly placed, so the this small low-frequency ILD may result from different shapes of the left and right ears of the HATS.) At symmetrical locations, these values were relatively equal in magnitude (less than 1.1 dB different). For VBAP, the average low-frequency ILDs closely matched (less than 0.8 dB different) with SC at LSP locations but diverged from the SC trend at locations between LSPs, particularly at L1 and R1. Average low-frequency ILDs for VBAP were symmetrical on the left and right (less than 1.8 dB different). For ambisonic sources, average low-frequency ILDs were close to SC across the horizontal plane (less than 2.2 dB different) except at L8, where they differed by 3.7 dB. Average high-frequency (4000-22050 Hz) ILDs for SC-presented sounds ranged from -12.09 dB to 12.25 dB, and were bilaterally symmetrical (less than 1.8 dB different), and trended monotonically from reference point L8 (–90°) through the midline (C) to R8 (+90°). High-frequency ILDs for VBAP matched SC values (less than 0.3 dB different), and were relatively symmetrical (less than 2.6 dB difference between left and right homologous locations). VBAP high-frequency ILD measurements diverged from interpolated SC data between LSPs, particularly at L1, R1, L7, and R7. Finally, high-frequency ILDs measured from ambisonic presentation were close to those measured for SC presentation between 69.09° and -50.41° azimuth locations (less than 1.9 dB different) and had

443 symmetrical magnitudes (less than 2.4 dB difference between left and right), but diverged
444 significantly and asymmetrically from SC data (and interpolations) at more lateral locations. Finally,
445 the standard deviation of the four ILD estimates did not exceed 0.07 dB at any location for any
446 reproduction method.

447 **C. Discussion**

448 Frequency responses to broadband sweeps were measured from the left and right ear canals of a
449 head and torso simulator (HATS), positioned in the AudioDome the same way as participants in
450 Experiment 1, for all reference points tested in Experiment 1. SC presentation at the frontal LSP
451 location ($\varphi = 0°$) produced responses in the left and right channels that are similar to each other and
452 to frontal HRTF magnitudes previously reported[2]. At all locations, within each channel, the
453 magnitude spectra of HRTF differ between ambisonic panning and SC presentation. Although the
454 power of ambisonically rendered sound sources is generally less than that of SC sources, the
455 difference in magnitude between the two reproduction methods is not homogenous across the
456 spectrum. In the high frequency range (4000-22050 Hz), the power of ambisonically rendered sound
457 is 1.5-8 dB below SC values for LSP-coincident reference points (the only exception is $\varphi = -90°$,
458 where high frequency power matches between reproduction methods). In the low frequency region
459 (0 – 4000 Hz) the power difference between the two methods was in the ±2.5 dB range, which is
460 overall lower than the power difference in the high-frequency region. Furthermore, the reproduction
461 method with higher power in this low-frequency band varies with the sound source location. (FIG
462 5.B)

463 Low-frequency (0-1800 Hz) ITDs for sources presented over SC followed a monotonic trend, f
464 from reference point L8 (–90°) through the midline (C) to R8 (+90°), within a range similar to
465 previous reports; for example, Feddersen et al. demonstrated ITDs between ±670 μs over this

range[2,25]. VBAP low-frequency ITDs matched these values at locations coincident with LSPs, as expected, but diverged from interpolated SC values at positions between LSPs. Low-frequency ITDs measured from ambisonically presented sounds were not only close to those of SC presented sounds, but they aligned better than did VBAP ITDs with the interpolated SC values at midpoint locations. For the higher frequency band (1800-22050 Hz), SC ITDs changed monotonically over the reference points from L8 to R8 as expected. VBAP and ambisonic ITDs diverged from interpolated SC values at locations between LSPs, and this divergence was more marked, and values were more asymmetrical between left and right, than for the low-frequency range. Fortunately, these higher-frequency ITDs are not relevant for perception and so these high-frequency ITD reproduction errors may not be perceptually salient. Thus our results indicate that neither VBAP nor ambisonic panning produce appreciable ITD errors (FIG 6, left column) and the ITDs accord well with those in the literature.

In the low-frequency region (0-4000 Hz), ILDs for SC-presented sounds are somewhat smaller than expected from published work. Feddersen et al[2,25] demonstrated ILDs between 0-20 dB for frequency subbands below 4000 Hz, with lower frequencies having smaller ILDs. Our low values might be due to averaging across the band. VBAP values matched those for SC presented sounds at LSP locations but diverged from the interpolated values at locations between LSPs. In contrast, although low-frequency ILDs for ambisonically presented sweeps were somewhat different from SC values at LSP locations, they were overall closer to the interpolated SC data across all locations. In the higher frequency range (4000-22050 Hz) ILD values measured from VBAP presentation were similar to those for SC presented sounds at LSP locations but still diverged from interpolated SC values between LSPs. High-frequency ILDs for ambisonically presented sweeps matched SC values at frontal locations (C, L1-4; R1-4), but diverged significantly from the SC trend at the most lateral locations (L7-8, R7-8). High frequency ILDs measured from ambisonically rendered sweeps differed

490 at homologous locations on the left and right. Given that ILD information used to localize sounds is
491 derived from high-frequency components of sounds these differences between ambisonic and SC
492 ILDs are likely to mislead human perceptual localization, especially when low-frequency ITDs are
493 not present to down-weight high-frequency ILDs[5,26](FIG 6, the right column). In general, binaural
494 localization information is largely preserved for frequencies below 4000 Hz in the AudioDome,
495 consistent with theoretical predictions.[7,14]ILD and HRTF cues at frequencies above 4000 may not be
496 correctly reproduced.

497 Finally, the spectral content, ITDs, and ILDs for sound sources presented ambisonically at the
498 most lateral locations ($\varphi = \pm 69.09°$ and $\varphi = \pm 90°$) differ on the right and left sides (for example,
499 look at the spectral power peaks at ~8 kHz in FIG 5.A). These unexpected asymmetrical
500 reconstruction errors are not observed when SC or VBAP reconstruction methods are used. The
501 asymmetries are most visible in the magnitude spectra. To ensure that this was not an artefact of the
502 recording setup or analysis method, the HATS was horizontally rotated 180° to face the back of the
503 AudioDome and the chirp sweeps were presented at $\varphi = \pm 90°$ locations with all three presentation
504 methods again. The asymmetries in the estimated magnitude spectra were reversed in the rotated
505 compared to the unrotated condition, indicating that the asymmetry results from the implementation
506 of ambisonic panning in the AudioDome. (See supplementary figure FIG.S. 1.) This implies that
507 there may be some asymmetry in the LSP or other hardware setup in the AudioDome, resulting in
508 somewhat different spectral errors at very lateral locations – beyond $\varphi = 69.09°$ - on the left and
509 right sides at frequencies over 4000 Hz. Interestingly, other studies that simulated or tested
510 ambisonic systems have also observed asymmetric reconstruction errors; Neal and Zahorik[13]
511 simulated SC, VBAP, and ambisonic reconstruction with different orders and estimated
512 reconstruction error and the potential impact on psychoacoustic localization cues (total level, ILD,
513 and ITD). Although not thoroughly discussed, they also report cue reconstruction errors that differ

514 between the left and right side in their simulations, especially for higher-order ambisonics. Other
515 examples can be found in Figures 9.11 and 9.13 in the sound-field review chapter of "Immersive
516 Sound" by Nicol[7] that illustrate ITD and ILD estimates across different locations on the horizontal
517 plane for a sound source at $\varphi = 0°$ rendered using 19th-order ambisonic system. Again, asymmetry of
518 reconstructed cues is observable, although not discussed and interpreted in the text. Ambisonic
519 panning might be very sensitive to implementation details such as LSP calibration, adjustment, etc.
520 Although it seems that reconstruction error asymmetry has not attracted much attention in the field,
521 it could be an issue when ambisonic technology is used for psychoacoustic experimentation;
522 erroneous conclusions about left-right asymmetry in human performance might be drawn when in
523 fact the source of the asymmetry is, at least in part, in the implementation of ambisonics.

524 **IV. EXPERIMENT 3**

525 During Experiment 1, participants anecdotally reported perceiving ambisonically presented
526 sound sources, particularly those at lateral locations, as elevated. This was surprising, since all sounds
527 were presented at locations in the horizontal plane. In Experiment 2, we determined that, at lateral
528 locations, spectral information above 4kHz was distorted when sounds were rendered using
529 ambisonic panning. Given that the spectral content in the 4-16 kHz range is important for sound
530 source elevation perception[21], these spectral differences might serve as unintended elevation cues. In
531 Experiment 3, we aim to provide empirical evidence for the anecdotal reports in Experiment 1 and
532 we hypothesize that they could be explained by the findings in Experiment 2. To achieve this goal,
533 we had an independent group of listeners compare the perceived elevation of one of a pair of noise
534 bursts, both presented at the same location: one reproduced with ambisonics and one presented
535 through a single channel. We tested the same reference points as in Experiment 1, and participants
536 were blind to the fact that sound sources were programmed to be collocated. Our hypothesis about

537 systematic errors in the reproduction of high-frequency information creating artefactual elevation
538 cues is investigated by comparing the degree to which lowpass, ambisonic sounds (i.e., without high-
539 frequency cues) are heard as elevated above their SC controls, compared to wideband, ambisonic
540 sources with high-frequency information, and thus artefactual elevation cues.

541 **A. Materials and Methods**

542 Twelve (5 female) young (aged 22-30 years), healthy, normally hearing (thresholds measured
543 audiometrically at both ears at octaves of 125 Hz up to 8 kHz; participants were excluded if their
544 threshold exceeded 25 dB HL at any tested frequency in either ear), right-handed listeners with no
545 history of neurological disorders completed the task. (Four listeners were excluded for the following
546 reasons: two reported concussion history, one did not pass the audiometry test, and one reported
547 during the debrief that they did not follow the eye fixation instructions.) Listeners had no experience
548 with AudioDome experiments. The experiment setup and participant positioning were similar to
549 Experiment 1 (except that no head tracking sensors were applied).

550 Participants completed an elevation discrimination task for reference points C, L4, R4, L6, R6,
551 L8, and R8 as listed in TABLE I. On each trial, participants were presented with two noise bursts
552 and were instructed to report the relative elevation of the second burst compared to the first one.
553 The trials were presented in two conditions: in the 'lowpass' condition both noise bursts contained
554 only frequency components below 4000 Hz (the boundary value for accurate acoustic environment
555 reproduction with ambisonic panning as explained in Experiment 2[7,14]), and in the 'wideband'
556 condition two wide-band noise bursts, with components between 0 and 22050 Hz, were presented.
557 In both conditions the noise bursts were presented from the same location: one was presented with
558 SC, and the other with 9th-order ambisonic panning. ($Ambi_{low}$ vs. SC_{low}, $Ambi_{wide}$ vs. SC_{wide} trials).
559 Twenty trials of each type were included for each reference point, with the SC-presented burst first
560 on half the trials. Another forty trials per reference point were tested in which both noise bursts

561 were presented with SC (SC_{low} vs. SC_{low}, SC_{wide} vs. SC_{wide} trials) as filler trials (they were not analyzed
562 later). To ensure participants stayed engaged, an additional 20 trials with large veridical elevation
563 differences, both bursts presented using SC, were included at each reference point. The azimuth
564 angle of sources in these trials was optimized to match the azimuth angle of the reference points,
565 therefore, based on the AudioDome's layout, the veridical elevation angle differences ranged
566 between 61° and 94°. Thus, 100 trials were tested per reference point; 700 trials total.

567 The stimuli were generated using MATLAB[22] with the same parameters as Experiment 1. The
568 lowpass noise bursts were originally generated as wideband (0-22050 Hz) and then, before the
569 attack-decay envelope was applied, were filtered with an FIR lowpass filter with a passband
570 frequency of 4000 Hz, a stopband of 4900 Hz, pass gain of 1 dB, stop gain of -80 dB, and a density
571 factor of 20, designed and generated with MATLAB's Signal Processing Toolbox[22,27]. Overall, 200
572 different noise bursts were generated and pseudo-randomly assigned to trials (the same noise burst
573 was assigned to both intervals of a single trial). The 700 trials were divided into five experimental
574 blocks, each with an equal number of trials of all trial types at all reference points. The order of trial
575 presentation was fully randomized. For each participant, stimulus assignment, trial order, and
576 presentation method order within Ambi vs. SC test trials were pseudorandomized such that all noise
577 bursts were presented an equal number of times at all reference points. .

578 As in Experiment 1, lights were extinguished, and a red LED was placed at the front, on the
579 horizontal plane at ear level for gaze fixation. Participants were instructed to report the relative
580 elevation of the second noise burst compared to the first one (above/below) via keypress on a small
581 keypad. At the beginning of the experiment, participants practiced the task on 14 trials in which
582 both bursts were presented single channel and there was a large veridical elevation displacement
583 between them (2 practice trials per reference point). They proceeded to the main trials once they
584 passed a 75% accuracy threshold in the practice block (the block was repeated after clarification of

585 the instructions if they performed with a lower accuracy). Feedback about accuracy was provided to
586 participants only during the practice block. Feedback about average reaction time (acceptable or too
587 slow, given the number of missed trials) and the number of response-missed trials was provided at
588 the end of each main block to keep participants focused on the task. The data were collected within
589 one session.

590 In the preprocessing stage, trials with no responses were removed from the data. Then, the
591 proportion of trials on which the ambisonically presented sound was perceived as elevated was
592 estimated for each participant, for all trial types, at each reference point.

593 **B. Results**

594 We predicted that ambisonically rendered sounds with frequency components over 4 kHz would
595 be heard as elevated compared to the same sound presented over a single channel on $Ambi_{wide}$ trials,
596 whereas ambisonically rendered sounds without such components would not (i.e., $Ambi_{low}$ trials). In
597 fact, in most cases the ambisonically presented sound was perceived to be below the SC sound. The
598 average proportion of trials on which the ambisonically presented sound was perceived as elevated,
599 as a function of reference point for both lowpass and wideband noise bursts, is, is illustrated in FIG
600 7.A ($Ambi_{low}$ and $Ambi_{wide}$). These values are reliably below the 50% value expected by chance at C
601 ($t_{low}(7) = -16.17$, $p < 10^{-6}$ & $t_{wide}(7) = -2.44$, $p < 0.05$), L4 ($t_{low}(7) = -20.11$, $p < 10^{-6}$ & $t_{wide}(7) = -$
602 5.23, $p < 0.05$), R4 ($t_{low}(7) = -6.45$, $p < 10^{-3}$ & $t_{wide}(7) = -15.56$, $p < 10^{-6}$), and L8 ($t_{low}(7) = -4.87$, p
603 < 0.01 & $t_{wide}(7) = -3.23$, $p < 0.05$). At R6 ($t(7) = -2.51$, $p < 0.05$) and R8 ($t(7) = -6.07$, $p < 10^{-3}$)
604 only the $Ambi_{low}$ proportion is below the chance level. All other values do not differ from the 50%
605 chance level ($p > 0.05$). It is possible that the timbral and power difference between ambisonically
606 rendered and SC rendered sounds was erroneously interpreted as an elevation difference, with
607 participants reporting the burst with less power (i.e., the ambisonically rendered burst) as lower in
608 elevation.

Even though subjects exhibited a response bias towards reporting the SC presented sound as higher, we could still ask whether there is a difference in the degree to which wideband ambisonically rendered sounds are heard as elevated, relative to the lowpass ambisonically rendered sounds (which presumably contain fewer misleading localization cues). We therefore tested the difference between $Ambi_{wide}$ –and $Ambi_{low}$. This value was computed individually for each person and is shown in FIG 7.B. This value will be positive if $Ambi_{wide}$ trials yield a greater proportion of elevation responses than do $Ambi_{low}$ trials, as predicted These values were indeed significantly positive at C ($t(7) = 2.06$, $p < 0.05$), L4 ($t(7) = 2.12$, $p < 0.05$), R6 ($t(7) = 6.25$, $p < 10^{-3}$), and R8 ($t(7) = 4.61$, $p < 10^{-2}$) reference points, ut not at R4, L6, and L8.

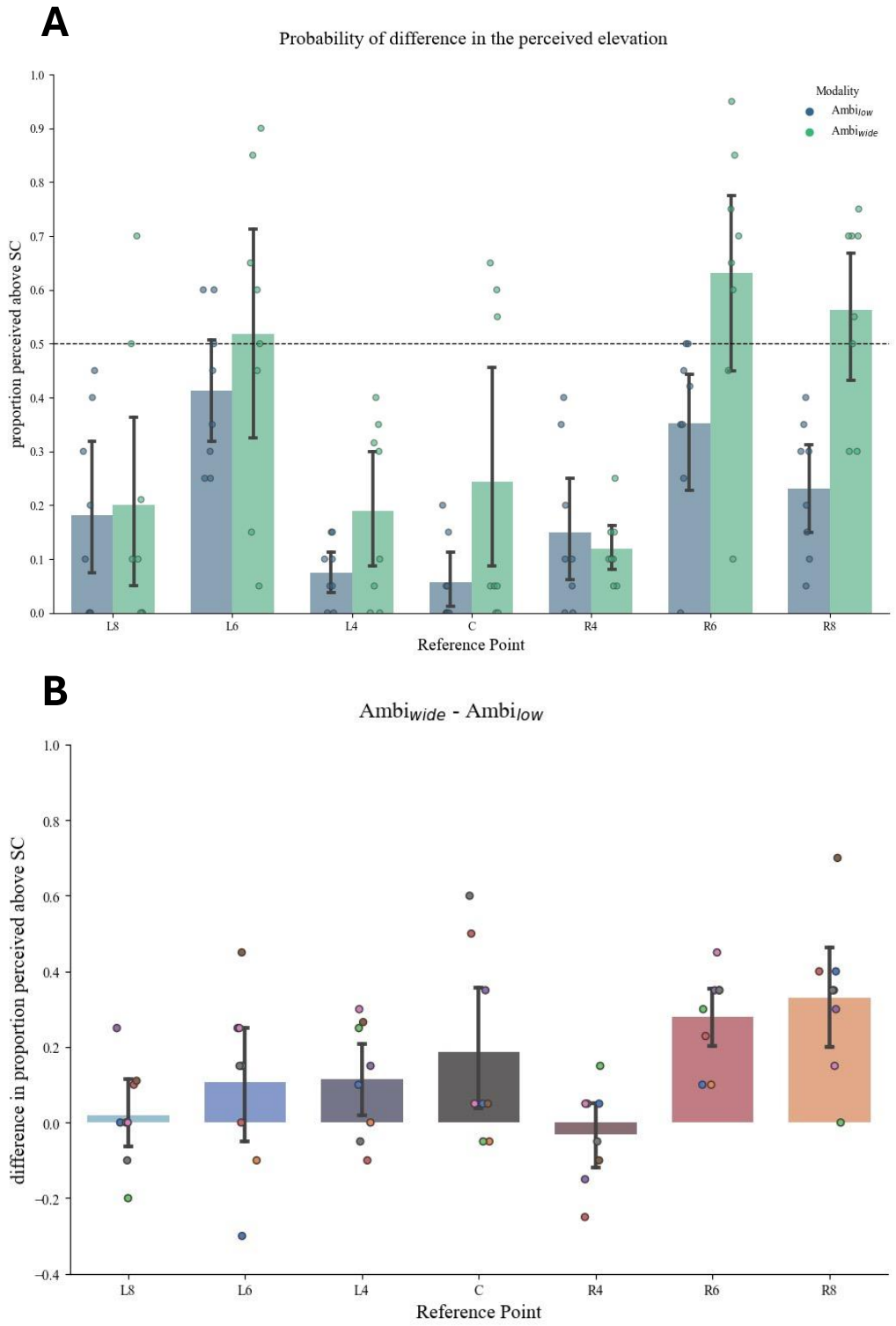


FIG. 7. (color online) A. The average proportion of ambisonically rendered sounds perceived as elevated, compared to the SC reference interval for each of the two experimental conditions,

621 $Ambi_{wide}$ and $Ambi_{low}$ B. The degree to which ambisonically rendered wideband noise bursts are
622 reported as elevated (relative to their SC reference) compared to ambisonically rendered lowpass
623 noise bursts (i.e., every set of two columns in panel A were subjected to the following arithmetic
624 operation: $Ambi_{wide} – Ambi_{low}$). Error bars in both panels represent the 95% confidence interval of
625 the grand mean. Dots show individual participant data. Except at R4 which the grand mean is
626 negative, all values are greater than zero, although only C, L4, R6 and R8 are significantly greater
627 than zero, demonstrating a reliably greater tendency for wideband ambisonically rendered sounds to
628 be perceived as elevated, compared to lowpass ambisonically rendered sounds, at these locations.
629 The lack of significance at the other locations could be due to a lack of power.

630 **C. Discussion**

631 In Experiment 3, we compared the perceived elevation of lowpass and wideband sounds,
632 ambisonically reproduced at horizontal reference points, relative to the same sounds presented from
633 a single loudspeaker at the same location. Rather surprisingly, the ambisonically rendered sounds
634 appeared to be generally perceived as below their SC controls. There is a perceptible timbral
635 difference between the ambisonic and SC rendered sources, and the ambisonic sources also have
636 lower power (as documented in Experiment 2). Participants may have interpreted these differences
637 in the ambisonic sounds to indicate a source below the horizontal plane.

638 Despite this general response bias, we observed that a greater proportion of ambisonic sounds
639 were perceived as elevated relative to the SC reference sound on wideband trials, compared to
640 lowpass trials, at almost all reference points (FIG 7.A). This consistent effect suggests that the high-
641 frequency components included in the wideband, but not lowpass, sounds were distorted by
642 ambisonic presentation, contributing to perceived elevation as predicted. However, there was also
643 more variability in response to wideband than lowpass trials. This is not surprising because each
644 participant has their own HRTF, so individuals will interpret the spurious spectral cues present in

645 the high-frequency components differently. Also, as illustrated in FIG 7.B, the effect of high-
646 frequency components in the perceived elevation of ambisonic sources seems to be somewhat
647 variable across different locations, being more pronounced at lateral locations on the right side. This
648 is consistent with our observations in Experiment 2 of larger reproduction errors on the right side,
649 compared to the left, possibly due to subtle asymmetries in the location or calibration of
650 loudspeakers. In this study, we assessed the fidelity of a 9th-order ambisonic loudspeaker array for
651 the reproduction of focal sound sources on the horizontal plane.

652 **V. CONCLUSIONS AND DISCUSSION**

653 In this study, we assessed the fidelity of a 9th-order ambisonic loudspeaker array for the
654 reproduction of focal sound sources on the horizontal plane. In Experiment 1, we quantified the
655 limits of human spatial acuity in the frontal half of the horizontal plane with wideband sound
656 sources presented with ambisonic panning. Values for the minimal audible angle (MAA), the just
657 noticeable difference in horizontal displacement ranged from 1.03 - 1.2°, for the horizontal midline
658 location, similar to previous findings.[1] In fact, MAAs were similar to the existing literature for
659 locations within the range of ±51° degrees azimuth[1], confirming that ambisonic panning in the
660 AudioDome results in sources that are so focal that any blurring is below the threshold for human
661 auditory spatial acuity. The recordings through a head-and-torso simulator in Experiment 2
662 demonstrated that ITD cues, estimated for ambisonic sources on the horizontal plane at frequencies
663 up to 1800 Hz, are not distorted.

664 Our system is theoretically restricted to accurately reproduce plane-waves up to 4 kHz for a
665 sphere of radius of 12 cm[7,14], and indeed we noticed distortions in the frequency region above 4
666 kHz, which altered the spectral profile in ways that might alter localization. Spectral magnitudes in
667 the range of ILDs were also affected, with left-right asymmetries that could create artefactual ILDs.

In Experiment 3, we showed that ambisonic reconstruction errors in the frequency region 4000-22050 Hz are associated with a greater tendency to perceive the rendered sources as elevated, compared to similar sounds without components in this frequency region. Overall, we showed that the AudioDome can reproduce focal sources reliably if high frequencies (above 4000 Hz) are omitted.

Based on these observations, for accurate localization it seems best to limit the reproduced spectral content in the AudioDome to components below 4000 Hz for accurate spatial reproduction. For naturalistic studies of speech listening in the presence of competing talkers and other sounds, inclusion of information above 4000 Hz is desirable and so SC reproduction may be preferable.

Additionally, we showed that ambisonic panning reproduces sources more homogenously and reliably than does VBAP, at locations between loudspeakers. We also observe a general difference of reproduction error of high frequency components of sound between left and right symmetrical locations (here, high frequency is above 1800 Hz for ITDs and above 4000 Hz for ILD and spectral cues); although this asymmetry might be only specific to our system, there is evidence of this phenomenon in other studies simulated ambisonic rendering[7,13]. While its basis might be related to the sensitivity of the ambisonic implementation to loudspeaker array calibration or positioning, the causes are unknown and future work is required to investigate to reveal the causes of this observation.

This project is the first study that aimed to utilize the AudioDome to test human perception, andand it is also a pioneer in conducting psychophysical experiments using a higher-order ambisonics system, demonstrating that sources rendered with a 9th order ambisonic system are sufficiently focal such that any blurring is below the threshold of human auditory spatial resolution.

Although there are issues with ambisonic rendering of components above 4 kHz, this research lays the groundwork for future experiments on naturalistic human auditory perception.

SUPPLEMENTARY MATERIAL

At the end of Experiment 2, the HATS was rotated horizontally 180° so we could measure the effect of physical orientation of the HATS on the recording content; it is possible that the asymmetrical reproduction errors are due to asymmetries in the HATS structure (the model's pinnae for left and right are not perfectly symmetrical as natural pinnae). If this hypothesis is true, then the response of the left and right channels should not change as we rotate the manikin and record the same sounds. However, as observed in FIG S.1, this is true for SC presentation, but for ambisonics, the response in the left channel transfers to the right ear and vice versa, showing that the recording is reflecting the true spectral content of the reproduced sounds at the recording location.

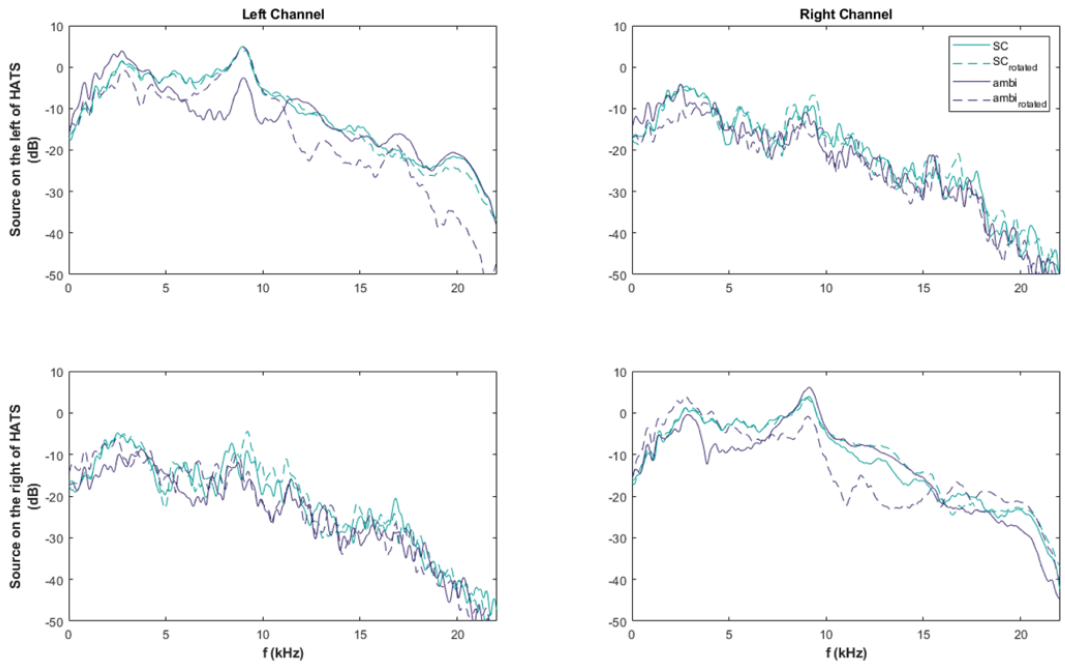


FIG.S. 1. (color online) The magnitude spectra of auditory responses to chirp sweeps presented from L8 and R8 (±90°) in both channels when the HATS is facing the center (φ = 0°) and when it is fully rotated to face the back of the AudioDome (φ = 180°).

ACKNOWLEDGMENTS

The authors would like to thank Dr. Derek Quinlan and Karsten Pravoslav Babin for their invaluable help and support with experiment hardware setup and programming. We thank the Canadian Foundation for Innovation (CFI) for funding the AudioDome, and the Natural Sciences and Engineering Research Council (NSERC) for research funding.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Ethics Approval

The experiments presented in this manuscript were approved by The Western University Non-Medical Research Ethics Board. Informed consent was obtained from all participants.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]A. W. Mills, "On the minimum audible angle," J. Acoust. Soc. Am. **30**, 237–246 (1958).

[2]J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, (MIT Press, 1997).

[3]C. J. Plack, *The sense of hearing*, (Third edition. | Abingdon, Oxon; New York, NY: Routledge, 2018., Routledge, 2018).

[4]Rayleigh, "XII. On our perception of sound direction," Lond. Edinb. Dublin Philos. Mag. J. Sci. **13**, 214–232 (1907).

[5]E. A. Macpherson, and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: the duplex theory of sound localization revisited," J. Acoust. Soc. Am. **111**, 2219–2236 (2002).

[6]H. Møller, "Fundamentals of binaural technology," , doi: 10.1016/0003-682X(92)90046-U. doi:10.1016/0003-682X(92)90046-U

[7]P. Roginska, *Immersive Sound, Chapter 9: Sound Field*, (Routledge, 2017), 1st Edition.

[8]S. Li, and J. Peissig, "Measurement of head-related transfer functions: A review," Appl. Sci. (Basel) **10**, 5014 (2020).

[9]V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," JAES **45**, 456–466 (1997).

[10]V. Pulkki, "Uniform spreading of amplitude panned virtual sources," in *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. WASPAA'99 (Cat. No.99TH8452)* (IEEE, 1999) pp. 187–190.

[11]F. Zotter, and M. Frank, *Ambisonics*, (Springer International Publishing, 2019).

[12]S. Bertet, J. Daniel, E. Parizet, and O. Warusfel, "Investigation on Localisation Accuracy for First and Higher Order Ambisonics Reproduced Sound Sources," Acta Acustica united with Acustica **99**, 642–657 (2013).

[13]M. T. Neal, and P. Zahorik, "The impact of head-related impulse response delay treatment strategy on psychoacoustic cue reconstruction errors from virtual loudspeaker arrays," J. Acoust. Soc. Am. **151**, 3729 (2022).

[14]D. B. Ward, and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," IEEE Trans. Speech Audio Process. **9**, 697–707 (2001).

[15]T. Z. Strybel, and K. Fujimoto, "Minimum audible angles in the horizontal and vertical planes: effects of stimulus onset asynchrony and burst duration," J. Acoust. Soc. Am. **108**, 3092–3095 (2000).

[16]G. A. Gescheider, "Psychophysics: The fundamentals.,"

[17]P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, et al., "SciPy 1.0: fundamental algorithms for scientific computing in Python," Nat. Methods **17**, 261–272 (2020).

[18]K. W. Vugrin, L. P. Swiler, R. M. Roberts, N. J. Stucky-Mack, and S. P. Sullivan, "Confidence region estimation techniques for nonlinear regression in groundwater flow: Three case studies," , doi: 10.1029/2005WR004804. doi:10.1029/2005WR004804

[19]The MathWorks Inc., *MATLAB 2022b*, (2022). Retrieved from https://www.mathworks.com

[20]V. Tabry, R. J. Zatorre, and P. Voss, "The influence of vision on sound localization abilities in both the horizontal and vertical planes," Front. Psychol. **4**, 932 (2013).

[21]J. Hebrank, and D. Wright, "Spectral cues used in the localization of sound sources on the median plane," J. Acoust. Soc. Am. **56**, 1829–1834 (1974).

[22]The MathWorks Inc., *MATLAB 2023b*, (2023). Retrieved from https://www.mathworks.com

[23]A. K. D. Nguyen, A. A. Simard-Meilleur, C. Berthiaume, R. Godbout, and L. Mottron, "Head circumference in Canadian male adults: Development of a normalized chart," Int. J. Morphol. **30**, 1474–1480 (2012).

[24]F. Brinkmann, A. Lindau, S. Weinzierl, M. Müller-Trapet, R. Opdam, M. Vorländer, and Others, "A high resolution and full-spherical head-related transfer function database for different head-above-torso orientations," J. Audio Eng. Soc. **65**, 841–848 (2017).

[25]W. E. Feddersen, T. T. Sandel, D. C. Teas, and L. A. Jeffress, "Localization of high-frequency tones," J. Acoust. Soc. Am. **29**, 988–991 (1957).

[26]M. L. Folkerts, and G. C. Stecker, "Spectral weighting functions for lateralization and localization of complex sound," J. Acoust. Soc. Am. **151**, 3409 (2022).

[27]The MathWorks Inc., *MATLAB Signal Processing Toolbox*, (2023). Retrieved from https://www.mathworks.com