Sound source localization with varying amount of visual information in virtual reality

Axel Ahrens*, Kasper Duemose Lund, Marton Marschall, Torsten Dau

Hearing Systems Group, Department of Electrical Engineering,
Technical University of Denmark, Kgs. Lyngby, Denmark

*Corresponding author
E-mail: aahr@elektro.dtu.dk

Abstract

To achieve accurate spatial auditory perception, subjects typically require personal head-related transfer functions (HRTFs) and the freedom for head movements. Loudspeaker-based virtual sound environments (VSEs) allow for realism without individualized measurements. To study audio-visual perception in realistic environments, the combination of head mounted displays (HMDs) and VSEs may be valuable. However, HMDs may affect the subjects' HRTFs and thus influence sound localization performance. Furthermore, due to limitations of the visual reproduction of the HMD, audio-visual perception might be influenced. Here, a sound localization experiment was conducted both with and without an HMD and with a varying amount of visual information provided to the subjects. Furthermore, interaural time and level difference errors (ITDs and ILDs) as well as spectral perturbations induced by the HMD were analyzed and compared to the perceptual localization data. The results showed a reduction of the localization accuracy when the subjects were wearing an HMD and when they were blindfolded. The induced changes of ITDs and ILDs were found to account for these errors. When visual information was provided, the localization error induced by the HMD was negligible. Presenting visual information of hand-location, room dimensions, source locations and pointing feedback on the HMD revealed similar effects as previously shown in real environments.

Introduction

Virtual environments (VE) and virtual reality (VR) systems enable the study of audio-visual perception in the laboratory with a higher degree of immersion than obtained with typical laboratory-based setups. Head-mounted displays (HMDs) may allow the realistic simulation of visual environments, and loudspeaker-based virtual sound environments (VSEs) can reproduce realistic acoustic environments while maintaining the subjects' own head-related transfer functions (HRTFs). Combining HMDs and loudspeaker-based VSEs could, therefore, be valuable for investigating perception in realistic scenarios.

To localize a sound source in the horizontal plane (azimuth) as well as in the vertical plane (elevation; see [1] for a review), three major cues are crucial: interaural time differences (ITDs), interaural level differences (ILDs), and monaural spectral cues generated by reflections from the body and the pinnae. An alteration of these cues can lead to a reduced localization accuracy [e.g. 2,3,4]. Wearing an HMD might thus alter the sound localization cues and reduce sound source localization accuracy.

Another factor that can affect people's sound source localization ability is visual information. The ventriloquism effect describes the capture of an acoustic stimulus by a visual stimulus [5,6], altering the perceived location of the acoustic stimulus. Maddox et al. [7] showed that the eye gaze modulates the localization accuracy of acoustic stimuli. They found an enhanced ITD and ILD discrimination performance when the eye gaze

was directed towards the source. Since HMDs have a reduced field-of-view relative to the human visual system, sound source localization abilities might also be affected due to reduced visual information. Furthermore, when having the room and the hand-location visible, the localization error has been shown to be lower than when subjects are blind-folded [8].

Modern proprietary VR systems have been shown to reproduce immersive visual environments and to provide reliable spatial tracking accuracy, both for the reproduction of virtual visual scenes as well as for headset- and controller-tracking [9,10]. However, Niehorster et al. [10] showed that when the tracking system of the HMD is lost, for example due to the user blocking the path between the tracking system and the HMD with their hands, the VR system fails to maintain the correct spatial location. Consequently, the calibrated location of the HMD within the room can be offset. Such offsets can be difficult to track in proprietary VR systems. To overcome this issue, a real-world to virtual-world calibration would be required.

The aim of the present study was to clarify how sound source localization ability is affected by a VR system, and in particular, how HMD-induced changes of the binaural cues and virtual visual information alter sound localization accuracy. In order to address this, a sound localization experiment was conducted in a loudspeaker environment with and without an HMD. For the localization task, a hand-pointing method was employed

utilizing commercially available hand-held VR controllers. To evaluate the accuracy of the pointing method, a visual localization experiment was conducted, where the subjects' task was to point to specific locations. Since previous studies showed that sound localization accuracy can be influenced by visual information [8,11], the amount of visual information provided to the subjects was varied in the VR environment. One condition included no visual information (blind-folded); another provided visual cues regarding the room dimensions and the hand-location; in a third condition, the subjects saw the loudspeaker locations and were additionally provided with a laser pointer for pointing on the perceived sound location. It was hypothesized that effects as observed in previous audio-visual localization experiments in real environments, may also be reflected in VR.

Methods

Subjects

Ten subjects (three female, seven male) with an average age of 24 years participated in the study. None of the subjects had seen the experimental room before. The subjects were blind-folded when they were guided into the room. They had normal audiometric thresholds equal to or below 20 dB hearing loss (HL) at the octave-band frequencies between 250 Hz and 8 kHz and self-reported normal or corrected vision. Nine of the ten subjects were right handed and were instructed to use their main hand to hold the controller. The data of the left-handed subject were mirrored on the median plane. For

each subject, the interpupillary distance was measured and the HMD was adjusted accordingly to ensure a clear binocular image and to minimize eye strain.

Acoustic reproduction method

The acoustic reproduction system consisted of 64 loudspeakers (KEF LS50, KEF, Maidstone, United Kingdom) housed in an anechoic chamber. The loudspeakers were arranged in a full sphere, at a distance of 2.4 m to the listening position, and driven by sonible d:24 (sonible GmbH, Graz, Austria) amplifiers. Twenty-seven of the loudspeakers in the frontal hemisphere were used in the present study. The loudspeakers were placed at three heights, at 0° (ear level) and ±28° elevation. Thirteen of the loudspeakers were at ear-level, distributed between -90° and +90° azimuth, with 15° separation. Seven loudspeakers were elevated by +28°, and seven by -28°, distributed between ±90° azimuth with 30° separation. All loudspeakers were equalized in level, delay and magnitude response as measured at the listening position. The loudspeakers were labelled using color coding (yellow, red and blue) to indicate elevation and numbers for azimuth locations (see Figure 1). The numbers ranged from one to thirteen, starting at -90° (left). The elevated loudspeakers at ±28° used only odd numbers, such that equal numerals were used for loudspeakers with the same azimuth angle.

Visual reproduction method

The real environment (RE), shown in Figure 1 (left panel), was replicated to ensure comparable visual information with and without the HMD (Figure 1, right panel). To present the VE, the HTC Vive system (HTC Corporation, New Taipei City, Taiwan) was used. This system consists of an HMD and two handheld controllers to interact with the VE. Three additional Vive Trackers (HTC Corporation, New Taipei City, Taiwan) were used for VE-to-RE calibration (see details below). The spatial position and rotation of all devices were tracked with the infrared ray-tracking system. Blender (Blender Foundation, Amsterdam, The Netherlands) and Unity3D (Unity Technologies, San Francisco, CA) with the SteamVR plugin (Valve Corporation, Bellevue, WA) were used to replicate and present the VE, respectively.

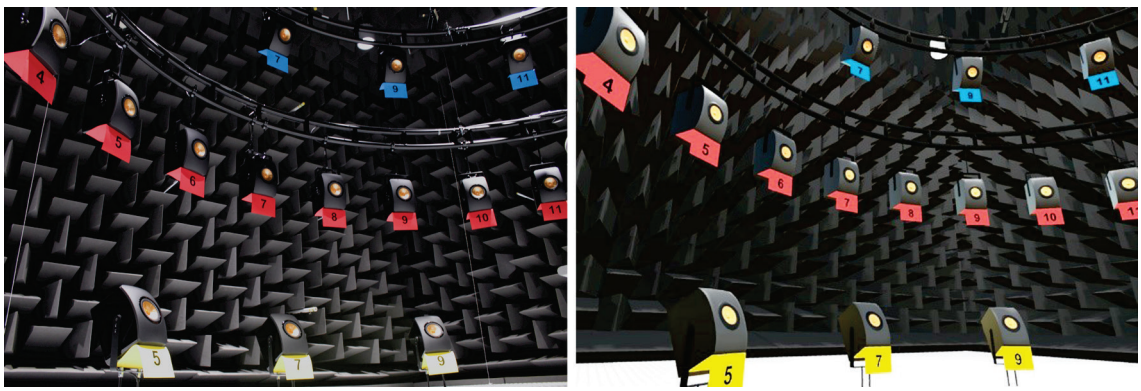

Figure 1. Photography (left) and screenshot (right) of the acoustic reproduction system in the real and in the virtual environment (RE and VE). The loudspeakers are numbered in azimuth and color coded in elevation.

For loudspeaker-based sound reproduction in VR and when the VE replicates a RE, it is crucial to ensure spatial alignment between the real and the virtual world. To calibrate the virtual world to the real world, the three Vive Trackers were positioned on top of the ear-level loudspeakers at 0° and ±45° azimuth. Discrepancies in the positions of the trackers in the RE and the VE were accounted for as follows:

1. The normal of the plane spanned by the three points corresponding to the positions of the trackers was calculated for the RE and for the VE and the difference in rotation between them was applied to the VE. This ensured the correct orientation of the VE.
2. To correctly position the VE, the difference between one tracker and the respective reference point in the VE was calculated and the VE was shifted accordingly. This resulted in an alignment of the RE and the VE at the chosen reference point.
3. The final rotation offset around the normal vector was corrected by calculating the angle difference of the vectors from the aligned reference point to an unaligned reference point in the VE and RE, i.e. to the known location of one of the other trackers and its virtual position.

After this procedure, the VE was aligned in both position and rotation relative to the RE. This method continuously accommodated for potential spatial discrepancies that might have occurred from tracking losses, as described by Niehorster et al. (2017). The system was recalibrated when either the tracker position error relative to the true position

exceeded 2 cm or when the HMD lost tracking. The maximum allowed positional offset of the reference points resulted in a worst-case rotation error of 3.2°.

Acoustic stimuli

The acoustic stimulus consisted of a pink noise burst with a duration of 240 ms and 20 ms tapered cosine ramps at the onset and offset. The noise burst was created in MATLAB (The Mathworks, Natick, MA) using a sampling frequency of 48 kHz. For each stimulus presentation, a new noise burst was created. The stimulus was presented at a sound pressure level (SPL) of 65 dB, and roved by values between ±3 dB, drawn from a uniform distribution. The short duration limits effect of head movements during the stimulus presentation [12] and the roving minimizes a spatial cue provided by directional loudness differences [13,14]. The subjects were asked to re-center their viewing direction before each stimulus representation, i.e., to face the 0° on-axis loudspeaker at ear level. The HMD rotation was logged in a subset of the conditions and for a subset of the subjects to evaluate if, on average, the viewing direction was centered at the time of the acoustic stimulus exposure. An initial azimuth rotation of the HMD of -1°±3° standard deviation was found.

Experimental conditions

Table 1 shows an overview of the eight experimental conditions considered in this study. The column ‘visual information’ shows the visual environment that was presented to the

subjects. The stimulus refers to the localization task, which was either visual localization (visual search) or sound localization. The last column indicates whether the HMD was worn or not. Each condition and each of the 27 source locations (see section Acoustic reproduction method) was presented five times to each of the subjects. The four blocks were presented in a fixed order from block I to block IV to control for the exposure to the real room and therefore the loudspeaker locations. The conditions were randomized within the blocks. No pre-experimental training was conducted to avoid any bias of the subjects with respect to a specific condition or regarding visual information.

Table 1. Overview over the conditions considered and grouped into blocks. Conditions were randomized within blocks. The conditions varied in available visual information, target stimulus and if the head-mounted display was worn.

Block	Visual information	Stimulus	HMD
I	Blind-folded	Acoustic	No
	Blind-folded	Acoustic	Yes
II	Virtual env., no loudspeaker (LS)	Acoustic	Yes
III	Virtual env., LS	Visual	Yes
	Real env.	Visual	No
	Real env.	Acoustic	No
	Virtual env., LS	Acoustic	Yes
IV	Virtual env., LS, laser pointer	Acoustic	Yes

The visual localization conditions in Block III were employed to investigate the baseline accuracy of the pointing method using the hand-held VR controller in the RE and the VE

181 (Figure 1). The subjects were instructed to point at a loudspeaker location shown either
182 on an iPad Air 2 (Apple Inc., Cupertino, CA) in the real environment or on a simulated
183 screen within the VE.

184 The two blind-folded conditions in Block I were used to examine whether, or to what
185 extent, simply wearing the HMD has an influence on sound source localization. The
186 reference localization accuracy was measured using the acoustic stimuli described above,
187 while subjects were blind-folded with a sleeping mask. To assess the effect of the HMD
188 frame on sound source localization, subjects wore the HMD on top of the sleeping mask.

189 To investigate the influence of the HMD on sound localization when visual information
190 regarding the loudspeaker positions is available, the sound localization accuracy was
191 evaluated in the real and in the virtual loudspeaker environments (Block III). The subjects
192 were informed that sounds could also come from positions in-between loudspeakers.
193 While the visual information regarding the loudspeaker positions was available in both
194 conditions, the VE provided reduced visual information. The field-of-view of the HTC
195 Vive is about 110° while the visual field of the human visual system is larger.
196 Furthermore, in the VE only the hand-held controller was simulated but not the arm,
197 which was inherently visible in the RE.

198 In addition to evaluating the pointing accuracy and the HMD-induced localization error,
199 the effect of varying the amount of visual information on sound localization in the VE

was investigated. In Block II, the experimental room (anechoic chamber) without the loudspeakers was rendered on the HMD and the subjects were asked to localize the acoustic stimuli as described for the previous conditions. The experiment thus included conditions with various degrees of visual information available to the subjects in the VR: no visual information, a depiction of the empty room including hand-location, and a depiction of the complete room including the locations of the sound sources.

To assess the role of visual feedback of the pointer location, in Block IV, a simulated laser pointer emerging from the hand-held controller was shown while the subjects completed the localization task in the VE with the room and the loudspeaker setup visible.

Pointing method

The controller of the VR system was used as a pointing device. The subjects were instructed to indicate the perceived stimulus location by stretching their arm with the wrist straight while holding the controller, in an attempt to minimize intra- and inter-subject variability in pointing. The pointing direction was defined by the intersection point of an invisible virtual ray originating at the tip of the controller extending the base of the device and an invisible virtual sphere with the origin at the listeners head position and the same radius as the loudspeakers (r=2.4m). The perceived position of the source was indicated with a button press using the index finger.

On each trigger button press, the PC running the VR system transmitted an Open Sound Control (OSC) message via User Datagram Protocol (UDP) over an Ethernet network to the audio processing PC. The audio processing PC subsequently presented the next acoustic or visual stimulus, with a delay of three seconds to allow the subject to re-center the viewing direction. With a responding OSC message, the audio processing PC permitted the reporting of the perceived location after the playback completed.

A virtual model of the controller was rendered in all conditions containing visual information in the HMD. Thus, the visual feedback in Blocks II and III was similar, independent of whether the HMD was worn or not. To standardize the pointing method for all audio-visual conditions, a direction marker, functioning as a visual pointing aid, was not provided in this study, except in the last condition (Block IV, Table 1), since a sufficiently comparable method was unfeasible in the real environment. Thus, the pointing method in Blocks II and III was similar to free-hand pointing.

Physical analysis

The effect of the HMD on the acoustic ear signals was analyzed from B&K Head and Torso Simulator (HATS; Type 4128-C; Brüel & Kjær A/S, Nærum, Denmark) measurements with and without the HTC Vive HMD. Binaural impulse responses were recorded with a 22-s long exponential sweep [15] in a frequency range from 60 Hz to 24 kHz and truncated to 128 samples (2.7 ms) to remove reflections from other

loudspeakers and objects in the room. Acoustic perturbations of the HMD on the frequency spectrum were analyzed by calculating the power in auditory filters between 200 Hz and 16 kHz with equivalent rectangular bandwidths [16] using the Auditory Modeling Toolbox [17]. The power in the auditory filters was averaged in three frequency regions from 200 Hz to 1 kHz, 1 to 5 kHz and 5 to 16 kHz.

Spectral differences (SD) were calculated as the absolute mean power differences of the three frequency regions, measured with and without the HMD. Interaural level differences (ILD) were determined in the same frequency region as the SD using the power differences at the output of the auditory filters between the left- and the right-ear signals. Interaural time differences (ITD) were calculated as the delay between the left- and right-ear signals. The delay of each impulse response was defined as the peak of the cross-correlation between the impulse response and its minimum-phase version [18].

Pointing bias

It was hypothesized that subjects might have a bias in pointing direction due to the shape of the hand-held controller, and because they had no knowledge on where the ‘invisible ray’ was emerging from the controller. A bias for each subject was therefore calculated in azimuth and elevation as the mean of all source locations in the two visual localization conditions (real and virtual). Individual responses were then corrected by the calculated

azimuth and elevation biases for all conditions except the condition with visual feedback of the pointer location (laser pointer condition).

Analysis of behavioral responses

Statistical analyses on the subject response errors were performed by fitting mixed-effects linear models to the azimuth (Az) and elevation (El) errors. The subject responses were corrected by a bias estimation due to the pointing method, as described above. Responses that were localized farther than 45° from the target location in either azimuth or elevation were treated as outliers. Of the 10800 subject responses, 0.29% were treated as outliers and discarded from the analysis.

Only the sources in the horizontal plane were considered in the statistical analyses of the azimuth localization errors. The azimuth stimulus location and the experimental condition (see Table 1) as well as their interaction were treated as fixed effects. Regarding the elevation error, the stimulus location in both azimuth (only azimuth directions that occurred in all elevation directions) and elevation, the experimental condition, as well as their interactions were treated as fixed effects. The influence of the subjects, the repetitions and their interaction were considered as random effects. The p-values were calculated based on likelihood-ratio tests for the random effects and on F-tests for the fixed effects [19]. Post-hoc analyses of within factor comparisons were performed using

estimated marginal means calculated from the mixed-effects linear models and using Bonferroni p-value adjustments [20].

Results

Pointing Bias

Figure 2 shows the pointing bias in azimuth (squares) and elevation (circles) for each subject calculated from the visual localization experiments. Regarding the azimuth bias, the subjects tended to point slightly too far to the left (-3.5° to -0.1°), except for subject S07, who had a slight positive (right) azimuthal bias of 1.3°. Overall, the average bias across subjects in azimuth was -1.6° (left). The only left-handed subject (S08) showed a similar bias as the other subjects. The bias in elevation angle (circle) was found to be higher than the azimuth bias for all subjects, with an average value of 19.0°. The subjects generally tended to point too high. The variance across subjects was between 12.8° and 28.6° and is likely to be related to the shape of the hand-held controller and the internal reference of each subject of where the “invisible ray” emerges from the controller. The responses of the subjects were corrected by the pointing bias for all conditions except the laser pointer condition.

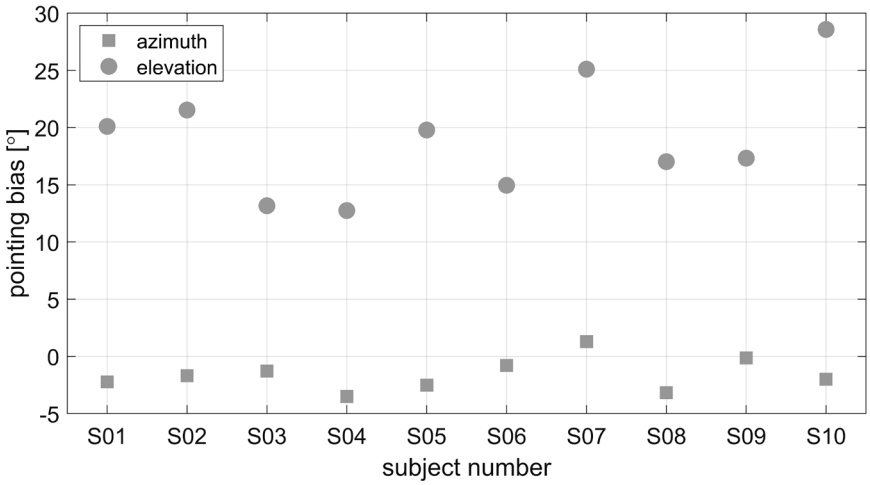

Figure 2. Pointing bias in azimuth (squares) and elevation (circles) for each subject. The bias was calculated as the mean error over all source locations in the two visual localization conditions. Negative angles indicate biases to the left and downwards for azimuth and elevation, respectively.

Spectral differences and interaural errors

Figure 3 shows the spectral difference (SD) of the left-ear signals obtained with and without the HMD on the B&K HATS for the azimuth source locations between -90° and +90° and for the elevation angles of -28° (downward triangles), 0° (diamonds) and 28° (upward triangles). For ipsilateral sources (negative azimuth angles), the SD was low for all frequency regions and all elevation angles. In the low-frequency region between 200 Hz and 1 kHz (dashed lines), the SD was also below 1 dB for contralateral sources (positive azimuth angles), independent of the elevation angle. The SD in the mid-frequency (dashed-dotted lines) and high-frequency (solid lines) regions was found to be

up to 6.3 dB for elevation angles at and above the horizontal plane (0°, 28°). For these elevations, the error was higher in the high-frequency region than in the mid frequency region, except at 60°. For mid- and high-frequency sources below the horizontal plane (i.e. an elevation angle of -28°), the SD was lower than for the other two elevation angles.

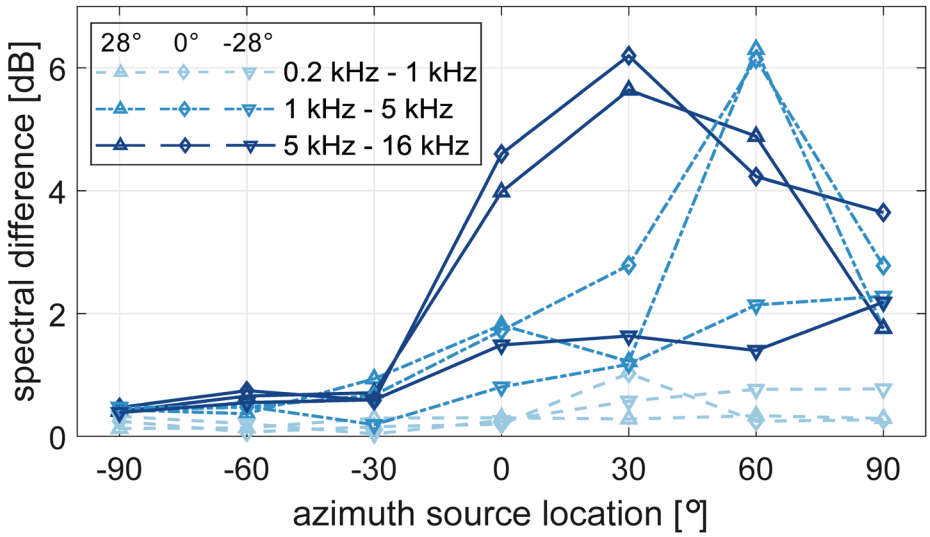

Figure 3. Spectral difference (SD) measured at the left ear of the B&K HATS with and without the HTC Vive. The SD was calculated in auditory bands and averaged over three frequency regions at low-, mid- and high frequencies.

Figure 4 shows the absolute errors in ILD (left panel) and ITD (right panel) induced by the HMD measured in the horizontal plane as a function of source azimuth angle. The ILD error in the low-frequency (dashed line) region was below 2 dB. In the mid- and high- frequency regions (dashed-dotted line and solid line), the error was lowest at the frontal source location (0° azimuth) and increased to 6 dB for lateral source angles. The

ITD error was below 1 sample (21 µs) for source angles between 0° and 30° and increased to 62.5 µs for the source angle of 75°. The ITD error was 0 µs at 90° azimuth angle.

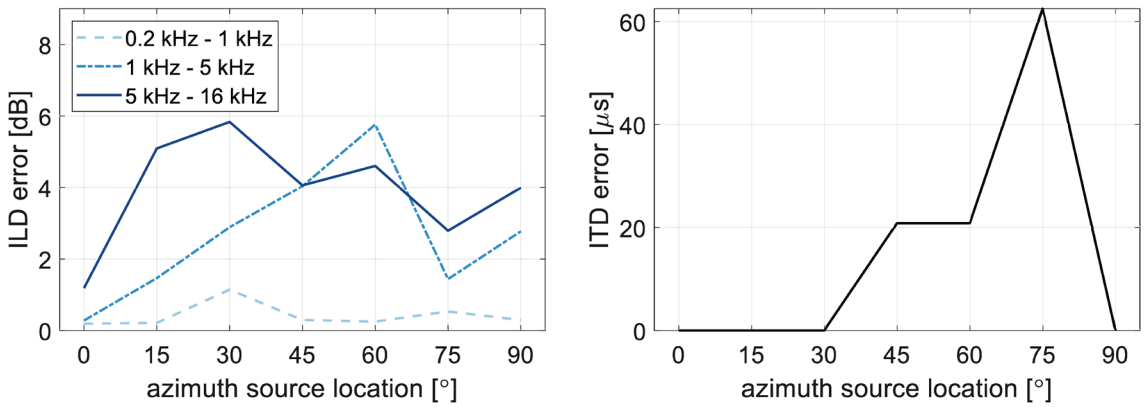

Figure 4. Absolute errors of interaural differences in level and time (ILD and ITD). The ILDs were calculated in auditory bands and averaged over three frequency regions at low-, mid- and high frequencies. The ITDs were calculated from the delays between the broadband binaural impulse responses (see Methods for details).

Pointing accuracy with VR controllers

Figure 5 shows the responses of the subjects in the two visual localization experiments in the RE and the VE. In both conditions, the subjects' task was to point to the center of the loudspeaker indicated on a screen. The filled black squares represent the 27 source locations. The small colored symbols represent the individual responses of the subjects and the large open black symbols indicate the mean responses, averaged across subjects and repetitions. The connecting lines between the target location and the mean responses indicate the localization error. The subjects generally pointed close to the correct

loudspeaker, whereby the precision of the responses was generally higher for azimuth than for elevation localization. The effect of the conditions (RE vs. VE) was not found to have an effect on the absolute azimuth error [$F(1,1263)=0.76$, $p=0.38$]. The azimuth location [$F(12,1263)=7.27$, $p<0.0001$], as well as the interaction between the conditions and the azimuth location [$F(12,1263)=1.83$, $p=0.0392$], were found to be significant. However, after correction for multiple comparisons, no difference between the RE and the VE was found for any azimuth location [$p>0.11$].

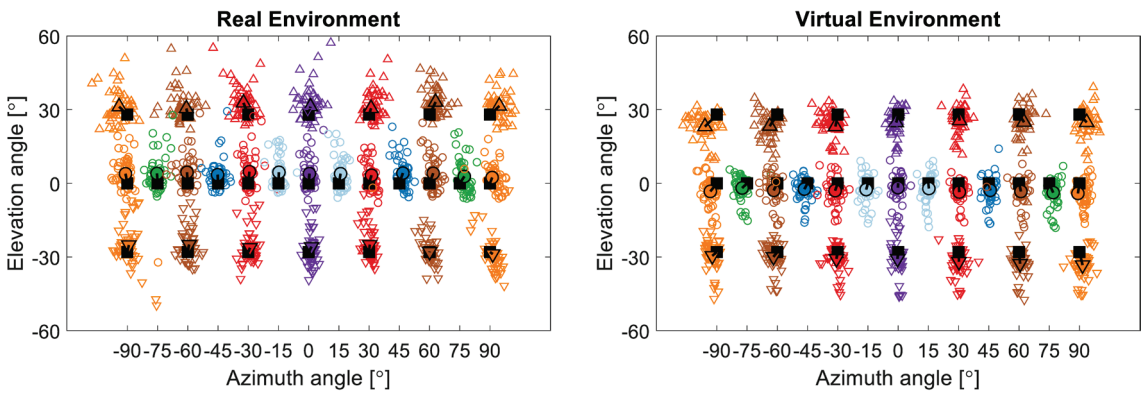

Figure 5. Response plot of the visual localization experiment for the real and virtual environment. The black squares represent the source locations. The small markers show the responses of the subjects and the large markers the mean response over subjects and repetitions. Negative angles represent sources to the left and downwards for azimuth and elevation, respectively.

The signed azimuth error was significantly affected by the conditions [$F(1,1263)=60.63$, $p<0.0001$], the azimuth source location [$F(12,1263)=28.02$, $p<0.0001$] and their interaction [$F(12,1263)=1.9$, $p=0.0305$]. The difference between the VE and the RE was

only significant for sources on the left at -90° [t(1263)=4.52, p=0.0001], -75° [t(1263)=4.26, p=0.0003] and -30° [t(1263)=3.71, p=0.0028]. Generally, the responses showed an overshoot, i.e. the sources on the right side showed a shift to the right, while the responses for sources on the left tended to show a shift to the left.

The analysis of the absolute elevation error showed that the conditions [F(1,2042)=9.77, p=0.0018] and the three-way interaction of condition, azimuth location and elevation location [F(12,1263)=2.02, p=0.0196] were significant. The two-way interaction of the conditions with the source locations in azimuth [F(6,2042)=2.21, p=0.0398] was significant, but not with the source locations in elevation [F(2,2042)=2.29, p=0.1]. The interaction between the source locations in azimuth and elevation was not significant [F(12,2042)=1.42, p=0.15] and also the source locations [Azimuth: F(6,2042)=0.99, p=0.43; Elevation: F(2,2042)=1.02, p=0.36] did not reveal significant effects.

The analysis of the signed elevation error showed significant contributions of the conditions [F(1,2069)=609.11, p<0.0001], the source azimuth [F(6,2069)=2.42, p=0.0245], the source elevation [F(2,2069)=7.85, p=0.0004] as well as the interactions between the source elevation and the conditions [F(2,2069)=9.51, p<0.0001] and both source locations [F(12,2069)=2.7, p=0.0013]. The interaction between the source azimuth and the conditions [F(6,2069)=0.75, p=0.61] and the three-way interaction [F(12,2069)=0.51, p=0.91] were not found to be significant. The difference between the

RE and the VE was significant at all three source elevations, at 0° elevation [t(2069)=15.48, p<0.0001], as well as below [t(2069)=10.76, p<0.0001] and above [t(2069)=16.52, p<0.0001] the horizontal plane. The signed elevation error was positive (upwards) in the RE and negative (downwards) in the VE, as indicated by the lines between the target markers (black squares) and the response markers (colored squares) in Figure 5. On average, the subjects pointed 6.5° higher in the RE than in the VE.

Influence of HMD on azimuth error

Figure 6 shows the absolute azimuth error as a function of the azimuth source locations. Negative angles represent sources on the left and positive angles indicate sources on the right of the subjects. The light grey boxes represent results for the condition where the subjects were blind-folded and the dark blue boxes show the results where the subjects were blind-folded and wore the HMD. The analysis of the linear mixed-effects model showed that the conditions [F(1,1265)=35.13, p<0.0001], the azimuth source locations [F(12,1265)=10.59, p<0.0001] and their interaction [F(12,1265)=3.07, p=0.0003] significantly affected the azimuth error.

The post-hoc analysis showed a significantly larger error when the subjects wore the HMD for source locations on the left side at -30° [t(1265)=-3.14, p=0.0225], -45° [t(1265)=-4.21, p=0.0003], at -60° [t(1265)=-4.36, p=0.0002] and at 45° [t(1265)=-3.15, p=0.0221] on the right side. At these angles, the increase in error with the HMD was

between 3.9° and 5.4°. Comparing the accuracy in localization on the left versus the right side showed no difference without the HMD [p>0.48] but a significantly larger error on the left side with the HMD between 15° and 60° [p<0.029].

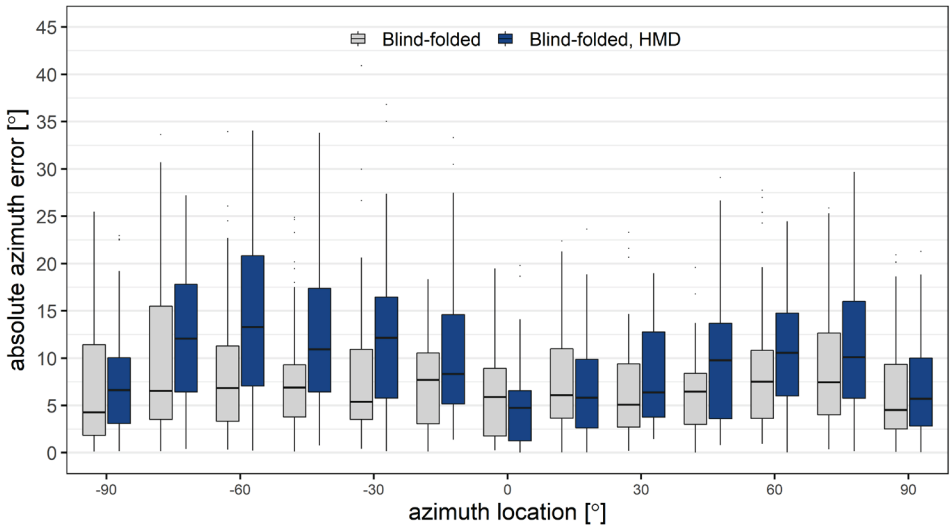

Figure 6. Azimuth error in degrees for acoustic localization for blind-folded subjects with (dark blue) and without (light grey) the head mounted display (HMD). The error is shown over the thirteen azimuth angles in the horizontal plane.

Influence of visual information on azimuth error

Figure 7 shows the absolute azimuth error as a function of the azimuth source locations for five conditions with varying visual information. The subjects were either blind-folded with the HMD (dark blue), were provided with information regarding the room dimensions and the hand position information (blue), could see the real room (grey), were

provided with the virtual version of the real room (light blue), or were provided with a laser pointer in the virtual room (white).

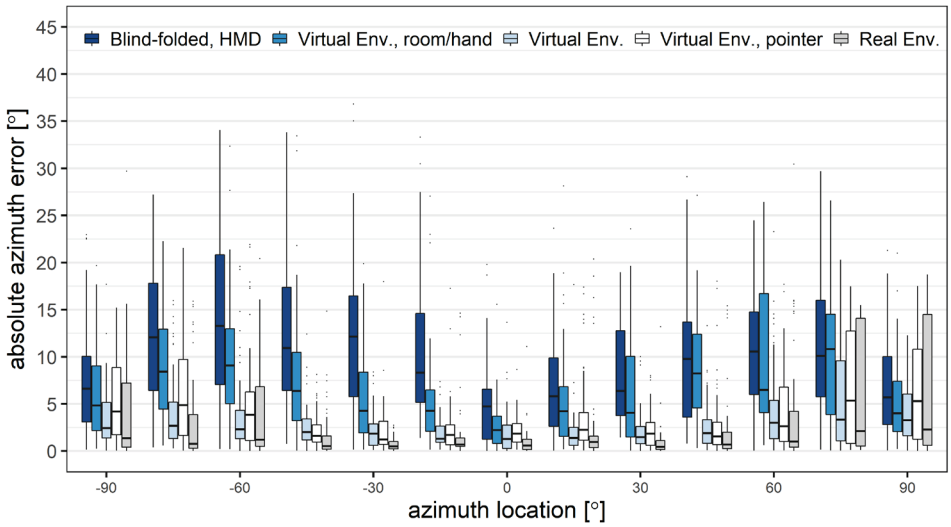

Figure 7. Azimuth error in degrees for acoustic localization with varying visual information in the virtual environment (VE) and the real environment (RE). The error is shown over the thirteen azimuth angles in the horizontal plane.

An analysis of the linear mixed-effects model showed that both main effects of azimuth source location [$F(12,3176)=23.17$, $p<0.0001$] and conditions [$F(4,3176)=223.88$, $p<0.0001$] as well as their interaction [$F(48,3176)=4.57$, $p<0.0001$] had a significant effect on the azimuth error. A post-hoc analysis was performed to investigate the effect of the room and hand position information, the loudspeaker locations, and the laser pointer for aided pointing on the perceived sound source.

A significant decrease in error was found when comparing the blind-folded HMD condition (dark blue) with the condition where the subjects had visual information of room and hand position (blue) at azimuth location between -15° and -75° [$p<0.039$]. No significant change in error was found at -90° [$p=1$], at the right side of the subjects [$p=1$] nor for the frontal source [$p=0.25$].

When also visual information of the loudspeaker locations was provided (light blue), the subjects generally pointed towards the correct source. When the loudspeaker locations were provided, the azimuth localization error decreased in comparison to the condition where only room and hand-location information were given. The reduction in error was significant for azimuth locations in the left quadrant between -15° and -60° [$p<0.03$] and in the right quadrant between 30° and 75° [$p<0.02$]. The lateral sources on the left [$p>0.34$] and on the right [$p=1$], as well as sources close to the midline at 0° and 15° [$p=1$] were not found significantly different in the two conditions with and without visual representations of the loudspeakers.

When the laser pointer was shown in the VE (light blue), the absolute azimuth error was not found to be different from the condition without the laser pointer (white) [$p=1$], except for the azimuth angle -75° [$p=0.0084$]. Comparing the VE without the laser pointer (light blue) and the RE (grey) showed no significant difference of the azimuth error at any of the source locations on the horizontal plane [$p>0.28$].

Influence of HMD on elevation error

Figure 8 shows the error in elevation as a function of the elevation target location. The results for the blind-folded conditions are shown with the HMD (light grey) and without the HMD (dark blue). The analysis of the linear mixed-effects model including the two conditions, azimuth and elevation locations, revealed significant main effects of conditions [$F(1,2049)=35.29$, $p<0.0001$] and source elevations [$F(2,2049)=8.28$, $p=0.0003$] but no effect of the azimuth locations [$F(6,2049)=1.98$, $p=0.0653$]. The interactions of the source locations with the conditions were not found to be significant. The mean increase in elevation error between the conditions with and without the HMD was 1.8°.

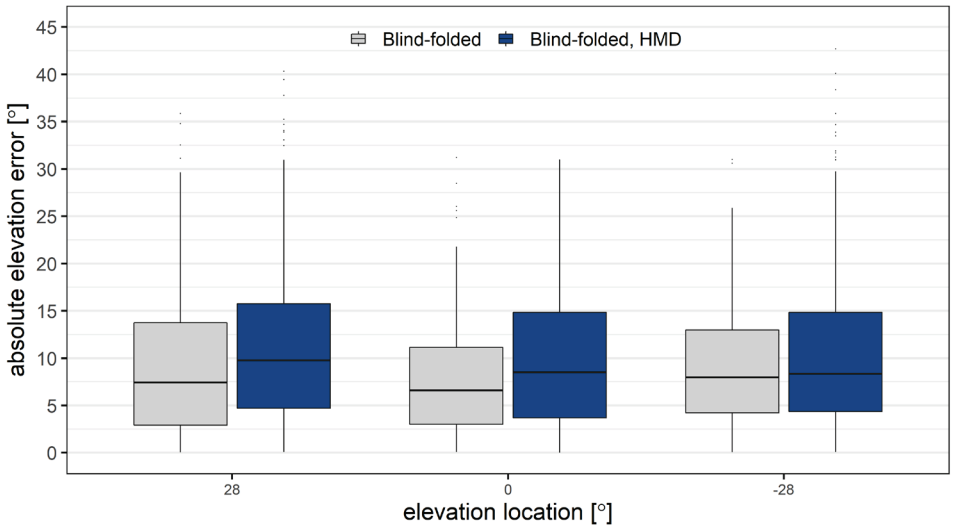

Figure 8. Elevation error in degrees for acoustic localization for blind-folded subjects with (dark blue) and without (light grey) the head mounted display (HMD). The error is shown over the three elevation angles and includes the sources in the horizontal plane.

Influence of visual information on elevation error

Figure 9 shows the absolute elevation error at the three source elevations for the five conditions with varying visual information. The statistical analysis showed significant main effects of the conditions [$F(4,5123)=289.44$, $p<0.0001$] and the azimuth locations [$F(6,5123)=9.47$, $p<0.0001$], but no effect of the elevation locations [$F(2,5123)=2.04$, $p=0.13$]. The interaction between the conditions and the locations was found to be significant for the elevation [$F(8,5123)=5.73$, $p<0.0001$] and non-significant for the azimuth [$F(24,5123)=1.43$, $p=0.0779$].

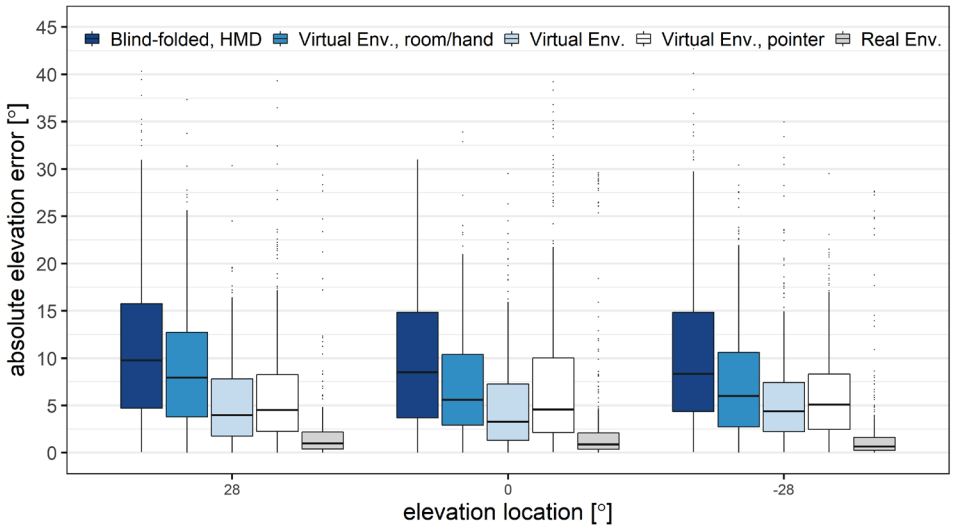

Figure 9. Elevation error in degrees for acoustic localization with varying visual information in the virtual environment (VE) and the real environment (RE). The error is shown over the three elevation angles and includes the sources in the horizontal plane.

The post-hoc analysis showed a significant drop in error between the blind-folded condition (dark blue) and the condition with visual room and hand-position information (blue) for all source elevations [$p<0.0001$]. The mean decrease of the error was 2.1° (from 10.2° to 8.1°). When the loudspeakers were visualized (light blue) in addition to the room and hand position (blue), the elevation error was found to further decrease significantly for the elevated sources [$p<0.0018$], but not for the sources in the horizontal plane [$p=1$]. When the laser pointer was employed (white), the absolute elevation error was lowest (2.4°) and significantly different from the error in the VE condition without the laser pointer (light blue) [$p<0.0001$]. The comparison of the elevation error in the RE (grey)

475 and the VE (light blue) revealed a significantly larger error for the sources at an elevation
476 angle of 0° [p<0.0001] but not for the sources above [p=0.16] and below [p=1] the
477 horizontal plane.

478 Discussion

479 Degraded sound localization with HMD

480 To localize a sound in azimuth direction, the auditory system relies on the interaural
481 disparities of time and level (ITD and ILD) [21]. The interaural disparities in the present
482 study were found to be larger when measured with the HMD than without the HMD,
483 consistent with results from the study of Gupta et al. [22]. Since Hafter & De Maio [23]
484 found ITD just-noticeable differences (JNDs) between 20 and 40 µs for lateral sources,
485 the differences in ITD of up to 60 µs induced by the HMD (see Figure 4) might be
486 perceptible. Likewise, the ILD changes of up to 6 dB induced by the HMD (see Figure 4)
487 were found to be above the JND of 2 dB [24,25].

488 The effect of the HMD on the perceived location of acoustic stimuli in an anechoic
489 environment was investigated in the blind-folded conditions. The error without the HMD
490 was comparable in azimuth and elevation to values reported in earlier studies [e.g. 26].
491 No difference in azimuth localization error with and without the HMD was found for
492 sources at or around the median plane, which is consistent with the small errors of ITDs

493 and ILDs induced by the HMD. However, for lateral sources, the azimuth error was larger
494 with the HMD than without the HMD, which is a consequence of the larger binaural
495 disparities caused by the HMD. However, the increase in localization error was larger on
496 the left side than on the right side. Thus, there seem to be additional factors contributing
497 to the localization error beyond the acoustical error induced by the HMD. Possibly, the
498 HMD represents an obstacle when pointing with the right hand to the left hemisphere,
499 resulting in a larger pointing error.

500 To localize sources in elevation, the auditory system mainly relies on spectral cues [27].
501 Recent studies investigated the effect of HMDs on HRTFs in a similar way as in the
502 current study and showed spectral perturbations up to about 6 dB at high frequencies
503 [22,28]. JNDs of such spectral perturbations have not yet been reported in the literature.
504 However, the results from the current study suggest that the HMD caused detectable
505 changes of the HRTF, since increased elevation errors were found.

506 Visual information influences sound localization in VR

507 In the virtual condition representing the simulated anechoic room without loudspeakers,
508 a reference frame was provided by the room and the subjects could see the pointer. The
509 sound localization error was found to be smaller with this visual information than in the
510 blind-folded condition with the HMD. The larger localization accuracy obtained in the
511 conditions when visual information about the hand position was provided might also have

been, at least partly, due to the visual information reflecting the room dimensions. Since the current study was not designed to distinguish between these two visual aspects, more experiments are needed to clarify their contributions. Tabry et al. [8] also observed lower errors of both azimuth and elevation sound localization in conditions similar to those in the current study whereby real visual information of the subjects' body and of the room were presented instead of virtual visual information. Thus, the amount of visual information provided in the current study may be considered to resemble those provided in a real environment. However, Tabry et al. [8] found substantially larger elevation errors than in the present study both in the condition with and without visual information. The smaller elevation errors found in the present study might be due to the smaller range of elevation source locations (-28°, 0°, 28°) than in the study of Tabry et al. [8], where the range of elevations was between -37.5° and 67.5°.

When the source locations were visible, the azimuth and elevation errors decreased by 3° and 1.5°, respectively, consistent with results obtained in real environments [29]. No improvement in accuracy was found for azimuth localization for frontal sources and for elevation localization of sources in the horizontal plane, since the auditory localization accuracy, even without visible source locations, is already high for frontal sources compared to those that are away from the midline and horizontal plane.

The additional information provided by the visual feedback from the laser pointer had only a negligible effect on the localization accuracy in azimuth, but a clear effect on the elevation accuracy. The elevation error was larger when no laser pointer was visible which might be partly due to the shape of the VR controller. The shape of the controller led to a biased pointing direction. The bias correction as described above (see Figure 2) was intended to reduce the influence of the controller. However, even though the subjects were asked to always hold the controller in the same way, the controller positioning might have varied leading to a larger pointing error when no visual feedback of the laser pointer was provided. Even though the effect of the laser pointer on the mean azimuth error was negligible, the variance of the subjects' responses decreased when the visual feedback of the pointing direction was available. Thus, the responses with the laser pointer become more reliable.

Comparing the localization error in the real and the virtual environments showed no differences in terms of the azimuth error. Even though the provided visual information was supposed to be the same in the two environments, some differences were unavoidable. While the peripheral visual field of the human visual system is above 200° in the horizontal plane [30], the field-of-view in the HMD is only about 110°. Thus, in the real environment, the whole range of loudspeakers (±90° azimuth) was visible when the gaze was towards the front, while with the HMD only sources between ±45° were within the field-of-view. Since it was shown that directing the eye gaze enhances ITD

550 and ILD discrimination [7], the effect of visual field reduction might also have affected
551 the localization error, especially for the lateral sources. Another difference between the
552 RE and the VE when the loudspeakers were visible was that the subjects could see their
553 arms in the RE, but not in the VE, where only the controller was visible. However, Van
554 Beers et al. [31] showed that the visual feedback of the arm does not seem to increase
555 visual object localization accuracy compared to the situation when only the finger is
556 visible.

557 Potential training effects

558 The subjects did not receive training before starting the experiment, but were introduced
559 to the task and the controller. Previous studies indicated that training can improve sound
560 localization accuracy [14,32,33]. Since the subjects could not be introduced to the visual
561 loudspeaker environment beforehand, no training was provided to avoid potential benefits
562 in certain conditions but not in others. It is possible that the localization performance of
563 the subjects improved throughout the course of the experiment. To minimize this effect,
564 the conditions were presented in a random order within the experimental blocks. The
565 random-order presentation avoided a bias within the blocks but could not exclude an
566 inter-block bias. Such inter-block bias was unavoidable because visual information
567 needed to be revealed to the subjects reflecting an increase of information content.

Implications for VR in hearing research

VR glasses in combination with audio reproduction techniques may allow for novel ways of conducting research in simulated realistic environments. While previous research typically involved simple audio-visual information, with VR, research can be conducted in ecologically more valid environments while maintaining controllability and reproducibility of the experiment. Even though the localization error increased with the HMD in the blind-folded condition, these errors may not be noticeable in realistic environments including visual information. This might also be the case for hearing-impaired subjects for whom sound localization accuracy is commonly reduced compared to normal-hearing subjects [34].

Conclusions

VR systems and loudspeaker-based audio reproduction allow for full immersion into an audio-visual scene. In the present study, sound source localization accuracy with an HMD providing a varying amount of visual information was investigated. A calibration system to align the real world and the virtual world was developed. Hand-pointing accuracy to visual targets was evaluated using commercially available controllers. The accuracy of the hand pointing to visual targets was found to be high in the azimuth direction, whereas a large bias was found in terms of elevation accuracy due to the shape of the controller. The sound localization experiment demonstrated a small detrimental effect of the HMD

on the sound localization accuracy. While the error induced by wearing the HMD was negligible in the frontal area, it was significant at more lateral sound source locations. The main source of this error was due to alterations of the HRTF induced by the HMD.

Generally, the sound localization accuracy was found to be higher when visual information was available than in the conditions without visual information. The lowest accuracy was found when the subjects were blind-folded and a significant improvement was found for both azimuth and elevation when room and hand position information were provided. When also the loudspeaker locations were visible, the error decreased further, whereas an additional laser pointer for pointing guidance did not lead to an improvement of azimuth localization but an improved elevation localization.

References

1. Blauert J. Spatial hearing: the psychophysics of human sound localization. MIT Press; 1997.

2. Hofman PM, Van Riswick JGA, Van Opstal AJ. Relearning Sound Localization with New Ears. Nat Neurosci. 1998;1: 417–421. doi:10.1523/JNEUROSCI.0850-05.2005

3. Shinn-Cunningham BG, Durlach NI, Held RM. Adapting to supernormal auditory localization cues. II. Constraints on adaptation of mean response. J

605 Acoust Soc Am. 1998;103: 3667–3676. doi:10.1121/1.423107

606 4. Shinn-Cunningham BG, Durlach NI, Held RM. Adapting to supernormal
607 auditory localization cues. I. Bias and resolution. J Acoust Soc Am. 1998;103:
608 3656–3666. doi:10.1121/1.423088

609 5. Howard IP, Templeton WB. Human spatial orientation. Human spatial
610 orientation. Oxford, England: John Wiley & Sons; 1966.

611 6. Recanzone GH. Interactions of auditory and visual stimuli in space and time.
612 Hear Res. 2009;258: 89–99. doi:10.1016/j.heares.2009.04.009

613 7. Maddox RK, Pospisil DA, Stecker GC, Lee AKC. Directing Eye Gaze Enhances
614 Auditory Spatial Cue Discrimination. Curr Biol. Elsevier Ltd; 2014;24: 748–752.
615 doi:10.1016/j.cub.2014.02.021

616 8. Tabry V, Zatorre RJ, Voss P. The influence of vision on sound localization
617 abilities in both the horizontal and vertical planes. Front Psychol. 2013;4: 1–7.
618 doi:10.3389/fpsyg.2013.00932

619 9. Borrego A, Latorre J, Alcañiz M, Llorens R. Comparison of Oculus Rift and
620 HTC Vive: Feasibility for Virtual Reality-Based Exploration, Navigation,
621 Exergaming, and Rehabilitation. Games Health J. 2018;7: g4h.2017.0114.
622 doi:10.1089/g4h.2017.0114

623 10. Niehorster DC, Li L, Lappe M. The Accuracy and Precision of Position and

Orientation Tracking in the HTC Vive Virtual Reality System for Scientific Research. Iperception. 2017;8: 204166951770820. doi:10.1177/2041669517708205

11. Dufour A, Després O, Pebayle T. Visual and auditory facilitation in auditory spatial localization. Vis cogn. 2002;9: 741–753. doi:10.1080/13506280042000250

12. Perrett S, Noble W. The effect of head rotations on vertical plane sound localization. J Acoust Soc Am. 1997;102: 2325–2332. doi:10.1121/1.419642

13. Musicant AD, Butler RA. The influence of pinnae-based spectral cues on sound localization. J Acoust Soc Am. 1984;75: 1195–1200. doi:10.1121/1.390770

14. Makous JC, Middlebrooks JC. Two-dimensional sound localization by human listeners. J Acoust Soc Am. 1990;87: 2188–2200. doi:10.1121/1.399186

15. Müller S, Massarani P. Transfer-Function Measurement with Sweeps. J Audio Eng Soc. 2001;49: 443–471. Available: http://www.aes.org/e-lib/browse.cfm?elib=10189

16. Glasberg BR, Moore BC. Derivation of auditory filter shapes from notched-noise data. Hear Res. 1990;47: 103–38. Available: http://www.ncbi.nlm.nih.gov/pubmed/2228789

17. Soendergaard P, Majdak P. The Auditory Modeling Toolbox. In: Blauert J,

editor. The Technology of Binaural Listening. Berlin, Heidelberg: Springer; 2013. pp. 33–56.

18. Nam J, Abel J, Iii JS. A method for estimating interaural time difference for binaural synthesis. Audio Eng Soc Conv 125. 2008; 0–6. Available: http://www.aes.org/e-lib/browse.cfm?elib=14763

19. Kuznetsova A, Christensen RHB, Bavay C, Brockhoff PB. Automated mixed ANOVA modeling of sensory and consumer data. Food Qual Prefer. Elsevier Ltd; 2014;40: 31–38. doi:10.1016/j.foodqual.2014.08.004

20. Lenth R V. Least-Squares Means: The R Package lsmeans. J Stat Softw. 2016;69. doi:10.18637/jss.v069.i01

21. Rayleigh, Lord. On our Perception of Sound Direction. Philos Mag. 1907;13: 214–232.

22. Gupta R, Ranjan R, He J, Gan W-S. Investigation of effect of VR/AR headgear on Head related transfer functions for natural listening. AES International Conference on Audio for Virtual and Augmented Reality. Redmond; 2018. Available: http://www.aes.org/e-lib/browse.cfm?elib=19697

23. Hafter ER, De Maio J. Difference thresholds for interaural delay. J Acoust Soc Am. 1975;57: 181–187. doi:10.1121/1.380412

24. Mills AW. Lateralization of High-Frequency Tones. J Acoust Soc Am. 1960;32:

132–134. doi:10.1121/1.1907864

25. Hafter ER, Dye RH, Neutzel JM, Aronow H. Difference thresholds for interaural intensity. J Acoust Soc Am. 1977;61: 829–834. doi:10.1121/1.381372

26. Oldfield SR, Parker SP a. Acuity of Sound Localization: A Topography of Auditory Space. I. Normal Hearing Conditions. Perception. 1984;13: 581–600. doi:10.1068/P130581

27. Hebrank J, Wright D. Spectral cues used in the localization of sound sources on the median plane. J Acoust Soc Am. 1974;56: 1829–1834. doi:10.1121/1.1903520

28. Genovese A, Zalles G, Reardon G, Roginska A. Acoustic perturbations in HRTFs measured on Mixed Reality Headsets. AES International Conference on Audio for Virtual and Augmented Reality. Redmond; 2018.

29. Shelton BR, Searle CL. The influence of vision on the absolute identification of sound-source position. Percept Psychophys. 1980;28: 589–596. doi:10.3758/BF03198830

30. Roenne H. Zur Theorie und Technik der Bjerrrumschen Gesichtsfelduntersuchung. Arch fuer Augenheilkd. 1915;78: 284–301.

31. Van Beers RJ, Sittig AC, Denier Van Der Gon JJ. Localization of a seen finger is based exclusively on proprioception and on vision of the finger. Exp Brain Res.

681 1999;125: 43–49. doi:10.1007/s002210050656

682 32. Carlile S, Leong P, Hyams S, Pralong D. The nature and distribution of errors in
683 the localization of sounds in humans [Internet]. Hearing Research. 1997. pp.
684 179–196. Available:
685 http://www.ncbi.nlm.nih.gov/pubmed/8988615%5Cnerrors_hr97.pdf

686 33. Majdak P, Goupell MJ, Laback B. 3-D localization of virtual sound sources:
687 Effects of visual environment, pointing method, and training. Attention,
688 Perception, Psychophys. 2010;72: 454–469. doi:10.3758/APP.72.2.454

689 34. Dobreva MS, O'Neill WE, Paige GD. Influence of aging on human sound
690 localization. J Neurophysiol. 2011;105: 2471–2486. doi:10.1152/jn.00951.2010

691