Audio-visual sound localization in virtual reality

THIRSA HUISMAN[1,*], TOBIAS PIECHOWIAK[2], TORSTEN DAU[1], AND EWEN MACDONALD[1]

[1] *Centre for Applied Hearing Research, Technical University of Denmark, DK-2800 Lyngby, Denmark*

[2] *GN Hearing, GN ReSound, Region Hovedsteden, Denmark*

Virtual reality (VR) can be a strong research tool in audio-visual (AV) experiments. It allows us to investigate AV integration in complex and realistic settings. Here, using a VR setup-up in combination with a loudspeaker array, 16 normal-hearing participants were tested on their sound localization abilities. The virtual environment consisted of a 1:1 model of the experimental environment except with the loudspeaker array replaced by a ring. This ring indicated the height, but not the position of the loudspeakers. The visual component of the stimuli consisted of a ball falling and then bouncing once on the ring after which it disappeared. As the ball collided with the ring, an impact sound was played from a loudspeaker. Participants were asked to indicate the apparent sound origin, for both congruent and incongruent visual and audio spatial positions ranging from -30 to 30 degrees. The VR visual stimuli in combination with real auditory stimuli were capable of inducing AV integration. The range of this integration extended, for several participants, over large ranges of AV disparity compared to some earlier studies.

INTRODUCTION

The integration of information from senses is a vital and well-studied topic, with often-cited studies dating back to 1950's. While in earlier studies, visual cues were thought to 'capture' the auditory cues, nowadays audio and visual cues are assumed to be integrated following bayesian causal inference (BCI). In BCI, the probability of a common cause is assessed. If there is no common cause, auditory and visual cues are processed independently. On the other hand, if there is a common cause, the cues are integrated optimally, such that they are weighted relative to their reliability. The actual perceived location of the integrated stimulus can then still vary depending on the decision-making strategies (Wozny *et al.*, 2010).

The assessment of a common cause is an important step in the model as audio-visual (AV) integration has been shown to be influenced by the timing and distance between the audio and visual stimuli, with increased disparity in either time or space reducing the probability of visual capture. Additionally, realism or 'compellingness' is often hypothesized to facilitate the ventriloquist effect. Despite this common assumption that more ecologically-valid stimuli might be integrated more strongly, experiments often use simpler stimuli, such as a light flash and a noise burst. The preference for these stimuli is most likely due to their ease of use. This is where virtual reality (VR) can be a valuable tool. VR allows us to simulate complex, ecologically-valid, and yet

*Corresponding author: thuis@dtu.dk

Proceedings of the International Symposium on Auditory and Audiological Research (Proc. ISAAR), Vol. 7: Auditory Learning in Biological and Artificial Systems, August 2019, Nyborg, Denmark. Edited by A. Kressner, J. Regev, J. C.-Dalsgaard, L. Tranebjærg, S. Santurette, and T. Dau. The Danavox Jubilee Foundation, 2019. ISSN: 2596-5522.

controlled scenarios. The current study was intended to function as a normal hearing baseline for a later comparison with hearing-impaired listeners, using VR as a tool to produce ecologically valid stimuli. We hypothesize that our VR baseline will match results of earlier studies on AV integration.

METHODS

Participants

Seven females and 9 males (average age 29.5 ± 13 years) were recruited from the DTU community. All had normal hearing thresholds and normal or corrected-to-normal vision. The procedure was approved by De Videnskabsetiske Komitéer for Region Hovedstaden (H-16036391) and all participants gave informed consent. The participants were compensated with an hourly rate of 122 DKK.

Apparatus

The experiment took place in the Audio-Visual-Immersion-Lab (AVIL) of the Technical University of Denmark. 5 loudspeakers were used to present the auditory cues. These loudspeakers were 2.4 m from the participant in an arc ranging from -30° to 30° azimuth with 15° separation between loudspeakers. Participants were seated in a height adjustable chair at the center. This chair was raised such that the participants' ears were at the height of the loudspeakers.

The visual cues were presented using an HTC VIVE VR headset. The virtual environment was a 1:1 model of the experimental room. However, the loudspeaker array was replaced by a gray ring. This ring, which was 5 cm in height, indicated the elevation and distance, but not the exact azimuth, of the loudspeakers. Only in the final condition of the experiment was the loudspeaker array shown. Just below the loudspeaker ring, at 0° azimuth, was a small white screen, with a visual angle (VA) of 10°. This was the focus point before and during the trials.

Participants could proceed through the experiment and record their judgements using a handheld HTC VIVE controller. In VR, a thin red rod was attached to the end of the controller so that it appeared to have a laser pointer. Participants pointed this "laser" at the location where they perceived the auditory stimuli and pressed a button to record their judgement.

Stimuli

The auditory stimulus was a 20 ms recording of the impact of a handball landing on a carpeted floor, presented at 65 dB peSPL. The visual stimulus consisted of an 8° VA ball. At the start of a trial, the ball appeared at a location above the ring, fell for half a second, bounced once on the ring and then, 20 ms after bouncing, disappeared. The audible impact of the ball on the loudspeaker ring was, on average, delayed by 105±15 ms relative to the visual impact of the ball.

Procedure

The experiment consisted of 4 blocks: unimodal audio, bimodal, unimodal visual and a pointing task. The blocks were presented in this fixed order. In the unimodal conditions and in the pointing task, 2 additional loudspeakers, at ± 45° azimuth, were included. These were not included in the bimodal conditions as these positions were near the limits of the field of view of the VR headset.

In the unimodal conditions, a stimulus was presented, randomly, at one of the 7 loudspeakers. Per position, the measurement was repeated 5 times for the auditory condition and 3 times for the visual condition. The AV block consisted of 322 trials. For each of the 5 loudspeakers used to present the auditory stimuli, visual stimuli were presented at the 7 loudspeaker positions and in a range of 30° around the loudspeaker position, using a 3° step size. This 30° range around the loudspeaker was limited to ± 45° for the speakers positioned at ± 30°, due, again, to limitations of the field of view of the VR headset. All combinations were repeated 3 times. In terms of AV separation, a maximum separation of ± 75° was tested, with the densest sampling occurring in the range of ± 15° disparity.

Trial

Participants were instructed to look forward at the focus point while the stimuli were presented. Once it was verified that their head was oriented towards the focus point, participants could press a button on the controller to start a trial. After the stimuli were presented, participants were asked to indicate where they heard the sound came from. In trials where there were no auditory stimuli, participants were asked instead to indicate where they saw the stimulus came from. Participants were allowed to move and look freely when pointing. Responses were restricted to the ring, such that the elevation and distance was fixed. Participants were, however, allowed to use the entire ring to answer, allowing for front-back confusions.

As eye movements have been shown to influence AV integration, an additional task was used to ensure that, at the moment of collision, participants were looking at the focus point straight ahead (rather than at the ball). As the ball collided with the ring, a letter appeared for 200 ms at the focus point. This letter was recognizable only when looking at the focus point. After performing the spatial localization task, a matrix of 16 different letters was presented and participants were asked to select the letter that had appeared during the trial. If an incorrect letter was selected, the trial was considered invalid and repeated again at a later random position.

The pointing condition was included to obtain the motor error in the pointing. Thus, it did not follow the above described structure. Instead, at the start of this condition, the ring was replaced by the model of the loudspeaker array. On each trial, participants were shown a number and were then instructed to point at the center of the loudspeaker labelled with that number. The loudspeakers were continuously visible and no auditory stimuli were presented in this condition. Hence, this condition estimated how well participants could point at a specific target.

RESULTS

All invalid trials were disregarded in the analysis. Due to a logging error in the pointing condition (where very fast responses could be logged as responses to the previous rather than current target), results with an error over 15 degrees azimuth (in this pointing condition only) were considered invalid.

First, the unimodal conditions were analyzed to predict the visual bias in the bimodal condition and to correct for localization biases found in earlier studies (e.g., Odegaard *et al.*, 2015; Ahrens *et al.*, 2019).

Figure 1 shows the average localization error of the unimodal conditions. As expected, the auditory localization error (max. 13.0° ± 17.6) and especially the variance was greatly increased, compared to the visual localization (max. 4.5° ± 2.9). The effects of the pointing method itself were very small, with the largest pointing error being about a single degree (max. 0.97° ± 0.4).

As in earlier studies (e.g., Odegaard *et al.*, 2015; Freeman *et al.*, 2018; Ahrens *et al.*, 2019), a bias in the unimodal conditions was found. For visual localization, a centralized bias, where visual stimuli are perceived more towards the center, was found (*t*-test, $p < 0.01$ for all non-zero locations). For auditory localization, an externalized bias can be seen for most, but not all locations (*t*-test, $p < 0.01$, for all but 15° and 30° azimuth). Surprisingly, a small externalizing bias, not centralized as might be expected in this visual task, was found ($p < 0.05$ for all paired Welch tests between adjacent angles).

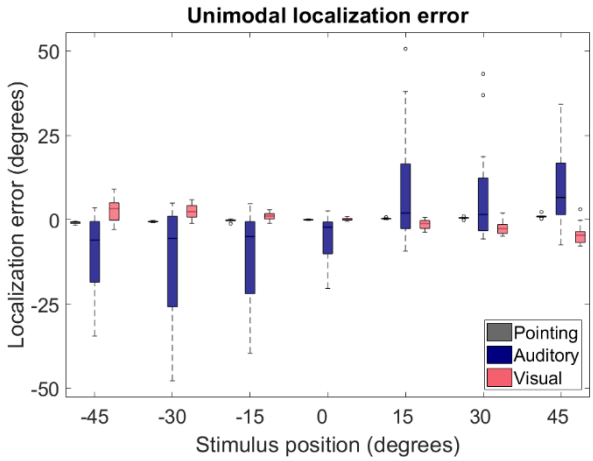

Figure 1: The localization error per modality for 7 stimulus positions. Stimuli presented at negative azimuths occurred in the left hemisphere and stimuli presented at positive azimuth occurred in the right hemisphere. Similarly, a negative localization error indicates that the stimulus was perceived leftwards of the stimulus positions, whereas a positive localization error indicates that the stimulus was perceived to the right of the stimulus position.

The bimodal data were clustered per participant using a Gaussian mixture models clustering (MATLAB, 2017b). The clustering was run with a maximum of 3 clusters to allow for audio, AV and visual clusters. The optimal number of clusters was then chosen using the BIC criteria (Schwarz, 1978). Figure 2 shows the cluster results for 3 participants. The clustering was run 20 times, after which the most prevalent clustering was used in the analysis.

As the task was to localize where the sound came from, we would expect that the localization error in Figure 2 would be around 0 (with some deviation due to the localization bias). Thus, we would expect clustering around the horizontal line in Figure 2. Indeed, at least one cluster with this property appeared for most participants. However, most participants also showed additional clusters, which were consistent with judgments being influenced by the position of the visual stimuli.

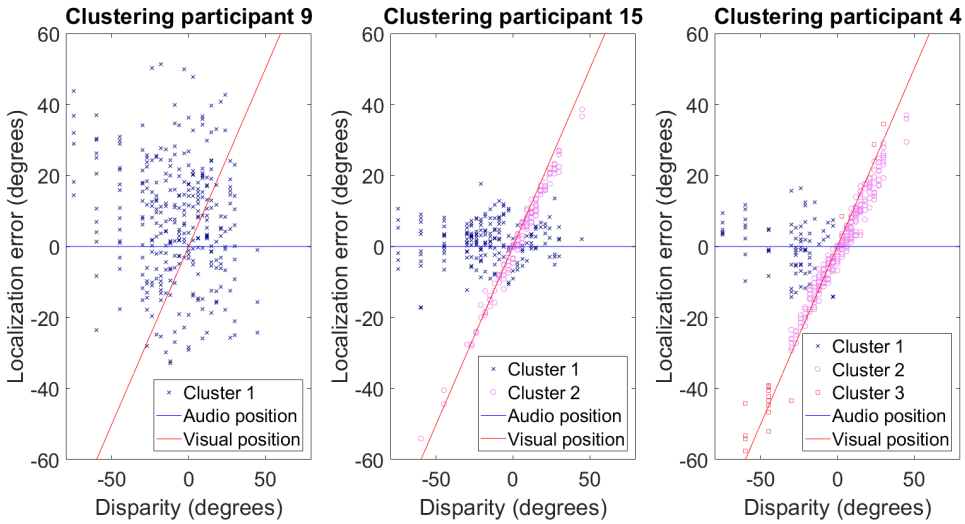

Figure 2: The panels show the cluster results for 3 participants, where, respectively, 1, 2 and 3 clusters were found. Each point is a single measurement, where the localization error is shown as a function of the disparity between the position of the auditory and visual stimuli. The results for the left hemisphere are mirrored such that a negative disparity indicates that the visual stimulus occurred closer to the midline. Additionally, the relative location of the auditory and visual stimulus is shown. The clusters are indicated with different symbols.

Through each cluster, a linear regression was fitted. The slope of each cluster was compared with 0, 1 and the predicted bias to see if it could be explained by either auditory localization, visual localization or AV localization. The predicted bias was based on the variance in the unimodal data. It weights the visual and auditory cues relative to the inverse of the localization variance for each modality, thereby predicting "optimal bimodal integration" (Alais and Burr, 2004). To ensure that the localization bias did not affect the categorization, the comparisons were also run with a correction for the auditory and visual bias. The results are shown in Table 1.

	Sub-categories	Number of clusters	Predicted visual bias	Visual bias	Range	
					Min	*Max*
Audio	*Average*	15	0.94	-0.02	-74.0	36.6
Visual*	*Average*	4	0.98	0.96	-67.5	41.3
Audio-visual	*Average*	16	0.92	0.57	-48.2	39.4
	Larger than 1	1/16	0.99	1.37	-27.0	30.0
	As predicted	2/16	0.87	0.86	-42.0	36.0
	Smaller than predicted	11/16	0.92	0.57	-43.9	39.3
	Smaller than 0	2/16	0.98	-0.11	-75.0	45.0

Table 1: 35 clusters were categorized as audio, visual* or AV based on the slope of the fitted linear regression curve. Audio had a slope that was consistent with auditory localization. Visual* was consistent both with visual and AV localization, but is assumed to be the result of visual only localization. All other clusters, where visual cues influenced, but did not dominate, auditory localization, were considered AV. The columns show, respectively, the number of clusters per category, the average predicted visual bias for these clusters, the average measured visual bias and the average range in degrees (min, max) over which these clusters occurred.

The distribution of the data points in each of the main categories of clusters listed in Table 1 is shown in Figure 3. As expected, the probability of AV responses is largest when the disparity is small. An asymmetry was found, in that the 'optimal' point for AV integration was shifted, such that the probability of integration is largest when the visual stimulus occurs more outwards compared to the auditory stimulus. Overall disparities with visual stimuli being presented from an eccentric position, relative to the position of the auditory stimulus (i.e. positive disparities), were more likely to result in AV integration.

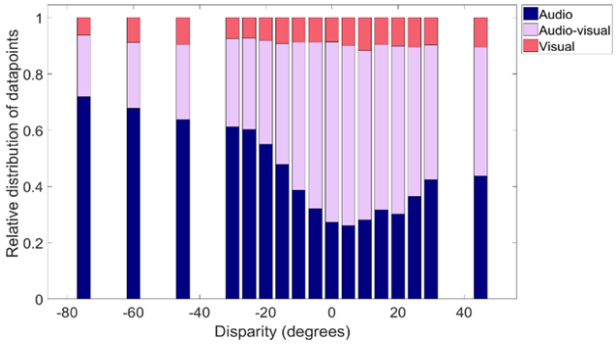

Figure 3: The relative distribution of data points in each cluster across the testing range, averaged over all participants.

DISCUSSION

While we found integration behavior similar to experiments using real world stimuli, there were some differences in our results. First, likely due to the delay between the audio and visual stimuli, we found many audio-only responses. For those, who did show integration behavior we found that the predicted bias generally overestimated the visual bias. This prediction relied on the unimodal results, where the audio-only condition showed a quite high variance on average, which would result in a very high predicted visual bias. Indeed, the predicted visual biases in this study were generally larger than the visual biases found by Battaglia *et al.* (2003). However, the average slope of AV clusters with an overestimated visual bias, is much lower than would be predicted by the visual biases found by Battaglia *et al.*

An alternative explanation might be in a combination of decision-making strategies. As mentioned earlier, the assumed decision-making strategy in most experiments with BCI is model selection. While the distribution of results in the present study could be explained with model selection, the stochastic manner in which the decision-making strategy switches within participants, sometimes from trial to trial, is more consistent with probability matching, which has been found to be the dominant strategy of most participants in a previous study (Wozny *et al.*, 2010). This probability matching by itself cannot explain a decrease in the visual bias. However, earlier studies (Battaglia *et al.*, 2003; Meijer *et al.*, 2019) also found a deviation from optimal integration. Their results showed an increased influence of visual position compared to what optimal integration would predict. They explained this by a model averaging strategy, where the AV and unimodal responses are weighted based on the probability of the underlying causal structure. As this experiment used mostly incongruent stimuli and also covered a wide range of incongruencies, the probabilities of the causal structures were strongly biased towards separate causes. Thus, model averaging could also explain the reduction in the visual influence observed here. As neither the probability matching nor model averaging strategy can account for the results by themselves, it appears that some of our participants combined different decision-making strategies.

Additionally, compared to recent studies (e.g., Bosen *et al.*, 2016), the AV results were found to extend over a surprisingly large range, with stimuli 75° (5 loudspeakers) apart being biased towards the visual stimulus. However, similar results have been found in much older studies (e.g., Jackson, 1953). Potentially, the use of more ecologically valid stimuli (which were also used in older studies) and/or the immersion of VR extends the range of spatial separation over which integration occurs.

These deviations from previous experiments, in particular the integration range, mean that the setup of the current experiment is of limited value for comparing the behaviour of normal-hearing and hearing-impaired listeners as the expected effects in the slope would be too small and the results of normal-hearing listeners already exhibit some visual influence over the full field of view available in VR. Potentially, when VR offers a larger field of view or when an alternative reproduction method is used for the visual stimuli, this method would be more applicable.

CONCLUSION

The VR ecologically valid stimuli used in the present study were capable of inducing integration showing that VR can be used for AV integration experiments. We found a deviation from the expected decision-making strategies. Some of the results here can only be explained by a combination of decision-making strategies (i.e., model averaging and probability matching). As integration occurred over a surprisingly large range, the current paradigm is not applicable for a comparison study between normal-hearing and hearing-impaired listeners with this setup.

ACKNOWLEDGEMENTS

This research was supported by the Centre for Applied Hearing research (CAHR) through a research consortium agreement with GN Resound, Oticon, and Widex.

REFERENCES

Ahrens, A., Lund, K. D., Marschall, M., and Dau, T. (**2019**). "Sound source localization with varying amount of visual information in virtual reality," PLoS ONE, **14**(3), 1–19. doi: 10.1371/journal.pone.0214603

Battaglia, P. W., Jacobs, R. A., and Aslin, R. N. (**2003**). "Bayesian integration of visual and auditory signals for spatial localization," J. Opt. Soc. Am. A., **20**(7), 1391. doi: 10.1364/JOSAA.20.001391

Bosen, A. K., Fleming, J. T., Brown, S. E., Allen, P. D., O'Neill, W. E., and Paige, G. D. (**2016**). "Comparison of congruence judgment and auditory localization tasks for assessing the spatial limits of visual capture," Biol. Cybern., **110**(6), 455–471. doi: 10.1007/s00422-016-0706-6

Freeman, L. C. A., Wood, K. C., and Bizley, J. K. (**2018**). "Multisensory stimuli improve relative localisation judgments compared to unisensory auditory or visual stimuli," J. Acoust. Soc. Am., **143**(6). doi: 10.1121/1.5042759

Jackson, C. V. (**1953**). "Visual Factors in Auditory Localization," Q. J. Exp. Psychol, **5**(2), 52–65. doi: 10.1080/17470215308416626

Meijer, D., Veselič, S., Calafiore, C., and Noppeney, U. (**2019**). "Integration of audiovisual spatial signals is not consistent with maximum likelihood estimation," Cortex, **119**, 74–88. doi: 10.1016/j.cortex.2019.03.026

Odegaard, B., Wozny, D. R., and Shams, L. (**2015**). "Biases in visual, auditory, and audiovisual perception of space," PLoS Comput. Biol., **11**(12), 1–23. doi: 10.1371/journal.pcbi.1004649

Schwarz, G. (**1978**). "Estimating the dimension of a model," Ann. Stat., **6**(2), 461–464.

Wozny, D. R., Beierholm, U. R., and Shams, L. (**2010**). "Probability matching as a computational strategy used in perception," PLoS Comput. Biol., **6**(8). doi: 10.1371/journal.pcbi.1000871