PLOS ONE

RESEARCH ARTICLE

Modeling the Perception of Audiovisual Distance: Bayesian Causal Inference and Other Models

Catarina Mendonça[1]*, Pietro Mandelli[2], Ville Pulkki[1]

1 Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland, **2** School of Industrial and Information Engineering, Polytechnic University of Milan, Milan, Italy

* Catarina.Mendonca@aalto.fi

Citation: Mendonça C, Mandelli P, Pulkki V (2016) Modeling the Perception of Audiovisual Distance: Bayesian Causal Inference and Other Models. PLoS ONE 11(12): e0165391. doi:10.1371/journal.pone.0165391

Editor: Christian Friedrich Altmann, Kyoto University, JAPAN

Received: April 7, 2016

Accepted: October 11, 2016

Published: December 13, 2016

Data Availability Statement: The experimental data is available in the Dryad Digital Repository at doi: 10.5061/dryad.r5gg0.

Funding: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant No 659114 and by the Academy of Finland, decision No 266239. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Studies of audiovisual perception of distance are rare. Here, visual and auditory cue interactions in distance are tested against several multisensory models, including a modified causal inference model. In this causal inference model predictions of estimate distributions are included. In our study, the audiovisual perception of distance was overall better explained by Bayesian causal inference than by other traditional models, such as sensory dominance and mandatory integration, and no interaction. Causal inference resolved with probability matching yielded the best fit to the data. Finally, we propose that sensory weights can also be estimated from causal inference. The analysis of the sensory weights allows us to obtain windows within which there is an interaction between the audiovisual stimuli. We find that the visual stimulus always contributes by more than 80% to the perception of visual distance. The visual stimulus also contributes by more than 50% to the perception of auditory distance, but only within a mobile window of interaction, which ranges from 1 to 4 m.

Introduction

Crossmodal interactions are often analyzed in light of the available multisensory perception theories of the time. For most of the twentieth century, these interactions were described by identifying which cue determined the multisensory percept. The terms sensory dominance, capture and ventriloquism were often used [1–6]. In the early 2000s the paradigm shifted away from the winner-takes-all perspective into a more probabilistic approach. Sensory interactions were expected to include weighing processes where the most reliable sensory cue contributed the most to the multisensory percept [7]. The Maximum Likelihood Estimation (MLE) model in particular, which assumes that this weighing process is statistically optimal, has been broadly tested and applied to a number of cue combination cases [8–14].

In recent years there has been a significant change in how multisensory interactions are described (see Fig 1 for an illustration of the multisensory models). No longer a unitary percept is expected to arise from multisensory stimulation [15]. It has been proposed that Bayesian causal inference mechanisms can explain human multisensory perception [16–20]. From bisensory stimulation perceivers can infer either one single causal event or two. The higher the

Competing Interests: The authors have declared that no competing interests exist.

temporal or spatial discrepancy, the more likely one is to infer two underlying events. It follows that cue integration is only expected to occur when both stimuli are perceived as stemming from the same physical event. These mechanisms have been shown to describe well the perception of audiovisual horizontal space [19, 20]. However, so far this model was only tested with generative models and using several free parameters. By free parameters we mean unknown values that the model does not predict. As a way of calculating the free parameters, the model was fit to the empirical data itself. Through the fitting procedure the values of the free parameters were obtained: they were the values that led to the best fit between the model and the data. This approach is valid, but there is a risk of overfitting and it is very computationally demanding. Here the causal inference model was tested with a new proposed approach. All model components were tested through numerical predictions and a generative model was avoided. We do so by proposing a new way of calculating the posteriors. This includes issuing predictions of distribution of estimates and of common causality.

The empirical study described in this manuscript analyzes and models the multisensory interactions in audiovisual distance perception. Psychophysical data were obtained in an experiment where visual and auditory stimuli of a person playing an organ were presented at

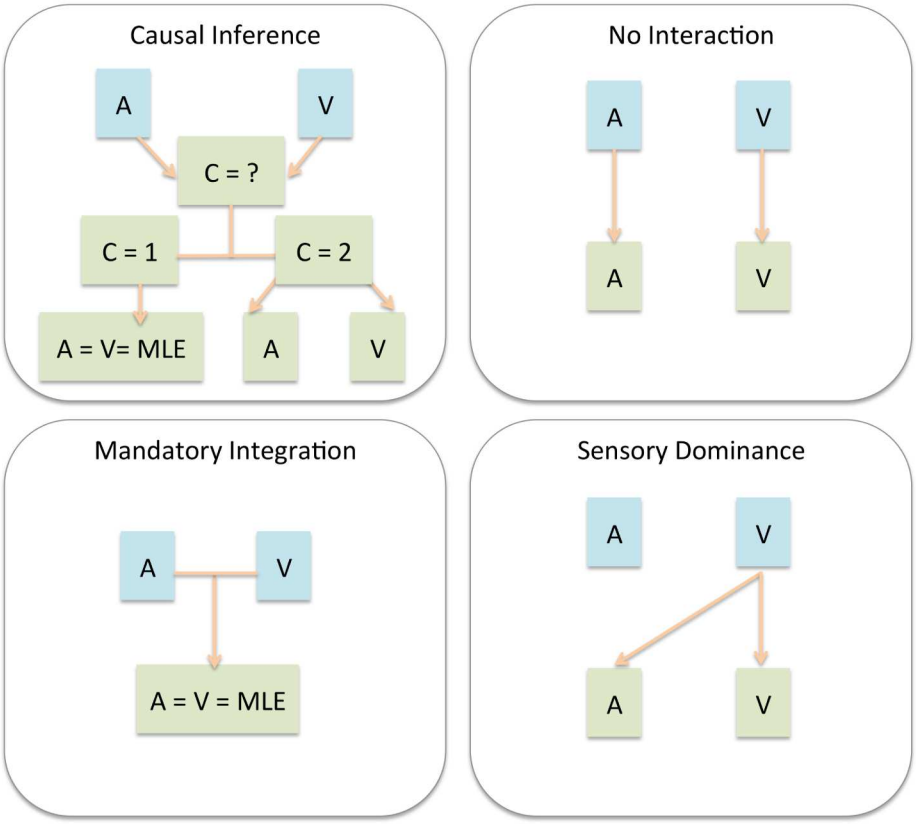

Fig 1. Representation of four models of multisensory interaction. Blue areas represent external events and green areas represent internal events. In *Causal Inference*, the first step is to find the likelihood of the stimuli having been caused by one or two sources. If one source is inferred, then *Mandatory Integration* takes place, which can be predicted by the MLE model. If two causes are inferred, then the estimates will be given by the *No Interaction* model. In the *No Interaction* model each sensory estimate in the multimodal condition can be approximated by the corresponding unisensory percept. In the *Sensory Dominance* model each sensory estimate in the multimodal condition can be approximated by the best unisensory percept. The models are described in detail in the section Multisensory Modeling.

doi:10.1371/journal.pone.0165391.g001

several distances inside a room. The bimodal trials consisted of several stimuli distance combinations. Subjects reported both the perceived visual and auditory distance of the stimuli. The above mentioned causal inference model is tested against other multisensory models. Sensory weights are calculated according to a new formula that accounts for causal inference. This allows for the description of cue interactions at several cue positions and discrepancies in space.

1 Multisensory Modeling

1.1 Causal Inference

Causal inference in multisensory perception assumes that multisensory stimuli can stem either from the same source or from separate sources. Further detail on causal inference in multisensory perception can be found in an article by [18]. This approach can be decomposed into four problems: 1) calculating the percept if a single cause is assumed; 2) calculating the percept if separate causes are assumed; 3) finding out the probability of common and separate causes; and 4) calculating the final percept accounting for all previous steps. These four problems have been previously solved [20] using a generative model with several free parameters. Here we solve them in a similar way, but without a generative model. A *causal inference* model is proposed where predictions of variance are added and the multisensory estimates are modeled assuming normal distributions.

1.1.1 Single Cause. When a single underlying event is inferred ($C = 1$), multisensory integration mechanisms can be predicted. We assume an unbiased perceiver whose estimates $\hat{s}_i$ can be well approximated by the underlying sensations x_i. The underlying sensations can be indirectly observed in the unisensory estimates. Note that while spatial prior biases may exist in the formation of estimates of unisensory signals from external stimuli (e.g. [20, 21]), no such biases are known between the estimate and the underlying sensation, or from the sensation under unimodal stimulation to the sensation under bimodal stimulation. If any bias were to be observed in the bimodal condition, it is hypothesized that it would be due to the concurrent stimulus, and not due to a change in the prior of the sensation itself. We also assume the sensations x_i to have a normal distribution with parameters $N(\mu, \sigma)$ and that added sensory noise is independent across modalities. In this case, the estimates $\hat{s}_i$ can be calculated using the Maximum Likelihood Estimation (MLE) model [7, 19]:

$$\hat{s}_{A,C=1} = \hat{s}_{V,C=1} = \frac{\frac{x_A}{\sigma_A^2} + \frac{x_V}{\sigma_V^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2}} \tag{1}$$

The benefit of multisensory integration is also observed in its predicted variance in which, when all assumptions are true, the inverse-variance-weighted estimate is also the minimum variance estimate of the stimulus property [7]:

$$\sigma^2_{A,C=1} = \sigma^2_{V,C=1} = \frac{\sigma_A^2 * \sigma_V^2}{\sigma_A^2 + \sigma_V^2} \tag{2}$$

1.1.2 Separate Causes. When two causes are inferred (C = 2), sensory estimates are not affected by the concurrent sensory stimulation. Therefore, they can be approximated by the corresponding unimodal sensations:

$$\hat{s}_{A,C=2} = x_A \text{ and } \hat{s}_{V,C=2} = x_V \tag{3}$$

In a similar way their variances correspond to the variance of the unimodal sensation. Therefore it is hypothesized that the variances of the estimates do not change from the unimodal to the bimodal condition, when two causes are inferred:

$$\sigma^2_{A,C=2} = \sigma^2_A \text{ and } \sigma^2_{V,C=2} = \sigma^2_V \tag{4}$$

1.1.3 Probability of Common Causes. In Eq (1) it is assumed that when there is a common cause sensory estimates are the same for both sensory modalities. Therefore, the causal probability can be indirectly observed by the similarity of the estimates $\hat{s}_i$. When the estimates are similar, a common cause can be inferred. A separate cause is inferred in the remaining cases:

$$\begin{aligned} p(C=1 \mid x_V, x_A) &= p(x_A = x_V) \\ &\textit{and} \\ p(C=2 \mid x_V, x_A) &= 1 - p(x_A = x_V) \end{aligned} \tag{5}$$

1.1.4 Calculating the Final Percept. The estimate of the probability of a common cause is rarely perfectly 0 or 1. When $p(C = 1)$ is any number between 0 and 1, it must be determined how the estimates from Eqs (1) and (3) are combined. We test three strategies proposed by [20].

In the model selection strategy one may simply choose the estimate from the most likely causal structure:

$$\hat{s}_{A|x_V,x_A} = \begin{cases} \hat{s}_{A,C=1} & \text{if } p(C=1) > 0.5 \\ \hat{s}_{A,C=2} & \text{if } p(C=1) < 0.5 \end{cases} \tag{6}$$

Let us assume an example where a given stimulus pair has $p(C = 1) = 0.3$. According to model selection, in our example, the estimates would always follow the most likely causal structure, which is $C = 2$. The estimates would therefore follow a normal distribution with mean and variance as observed in the unimodal condition.

In the model averaging strategy subjects have access to both independent estimates and provide a combined estimate. The final estimate is a linear weighted average of the estimates from both causal structures.

$$\hat{s}_{A|x_V,x_A} = p(C=1 \mid x_V, x_A) * \hat{s}_{A,C=1} + p(C=2 \mid x_V, x_A) * \hat{s}_{A,C=2} \tag{7}$$

In our example, the final auditory estimate would therefore be the sum of 30% of $\hat{S}_{A,C=1}$ with 70% of $\hat{S}_{A,C=2}$. To test this strategy, and in the absence of a better prediction from previous research, we also hypothesize that the variance of this estimate is a linear weighted average of the predicted variances for each causal structure.

$$\sigma^2_{A|x_V,x_A} = p(C=1 \mid x_V, x_A) * \sigma^2_{A,C=1} + p(C=2 \mid x_V, x_A) * \sigma^2_{A,C=2} \tag{8}$$

A third strategy, probability matching, assumes that estimates vary from trial to trial, following either $\hat{s}_{A,C=1}$ or $\hat{s}_{A,C=2}$. Each estimate type occurs proportionally as many times as its causal probability:

$$\hat{s}_A = \begin{cases} \hat{s}_{A,C=1} & \text{if } p(C=1 \mid x_V, x_A) > \zeta \\ \hat{s}_{A,C=2} & \text{if } p(C=1 \mid x_V, x_A) < \zeta \end{cases} \tag{9}$$

where ζ is sampled randomly from the uniform distribution [0:1]. Returning to our example, in practice, the estimates will follow the mean and variance of $p(C = 1)$ in 30% of the trials and of $p(C = 2)$ in the remaining trials. It follows that the response distribution according to this strategy is composed of two gaussian distributions, each with the relative size of each causal probability.

1.2 Models without causal inference

All the strategies described above assume causal inference. Additional models that do not assume causal inference were also tested. This testing aimed at establishing if causal inference is likely to take place in the perception of visual and auditory distance. Three alternative explanatory models were tested. *Sensory dominance* was tested by assuming that the most reliable cue always takes over. Therefore, the sensory estimates $\hat{s}_i$ in the multisensory condition can be approximated by the perceived distance x_i in the unimodal condition that had the lowest variance, as given by

$$\hat{s}_{A|x_V,x_A} = \hat{s}_{V|x_V,x_A} = \begin{cases} x_A & \text{if } \sigma_V^2 > \sigma_A^2 \\ x_V & \text{if } \sigma_V^2 < \sigma_A^2 \end{cases} \tag{10}$$

and in a similar manner the distribution of the estimates in the multisensory condition corresponds to the smallest distribution of the estimates of the unimodal sensation. *Mandatory integration* was tested by assuming that all estimates follow the linear inverse variance weighting rule. Therefore, the estimates in the multisensory condition are given by Eq (1), and their distribution is given by Eq (2). Finally, *no interaction* was also tested, where all estimates and distributions were given directly by the respective unimodal sensation, as described in Eqs (3) and (4).

1.3 Sensory Weights

Finally, we aimed at quantifying the relative contribution of each sensory cue by applying the principles of causal inference. In the conventional MLE model [7] it is assumed that the sensory weight is given by the normalized inverse of the variance of the unisensory estimate:

$$w_A = \frac{1/\sigma_A^2}{1/\sigma_A^2 + 1/\sigma_V^2} \tag{11}$$

To calculate the sensory weights accounting for causal inference one must calculate separately the weights for each causal structure. In the case of dual causality, it is assumed that the cues do not interact. Therefore, the weight of the auditory stimulus on the auditory distance estimate equals 1. In the case of a common causality, one may assume the weights as predicted by the MLE. Both in the probability matching and in the model averaging strategy, the average sensory weights correspond to the combination of the weights of perceived common (p(C = 1)) and separate (p(C = 2)) causes:

$$w_A = w_{A,C=2} * p(C = 2|x_V, x_A) + w_{A,C=1} * p(C = 1|x_V, x_A) \tag{12}$$

Therefore, the sensory weights can be calculated as follows:

$$w_A = p(C = 2|x_V, x_A) + p(C = 1|x_V, x_A) * \frac{\sigma_V^2}{\sigma_A^2 + \sigma_V^2} \tag{13}$$

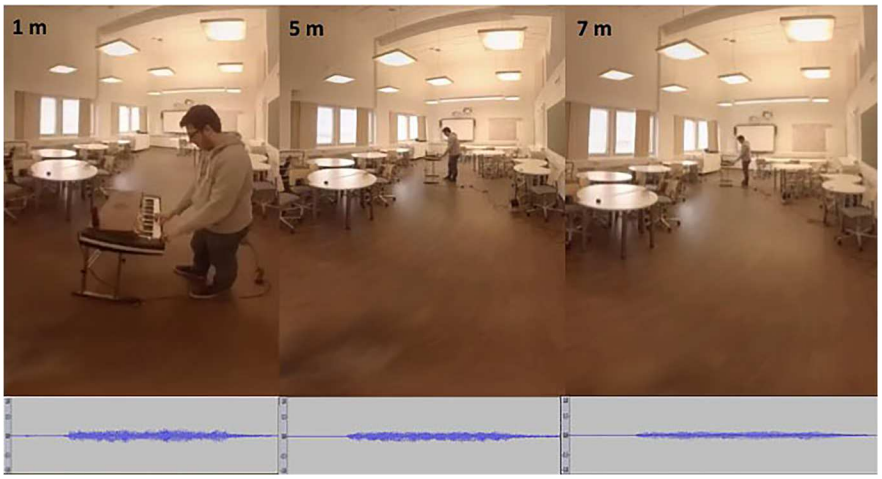

Fig 2. Stimuli at 1, 5 and 7 m in distance.

doi:10.1371/journal.pone.0165391.g002

and

$$w_V = p(C = 1|x_V, x_A) * \frac{\sigma_A^2}{\sigma_A^2 + \sigma_V^2} \quad (14)$$

where w_i is the weight of each sensory modality on a given auditory estimate.

2 Materials and Methods

2.1 Ethics Statement

The experiment followed the policies on human subjects research as described in the Declaration of Helsinki. Participants provided written informed consent. The experimental protocol was approved by the Aalto University Ethics Committee. The individual in Fig 2 of this manuscript is an author and has given written informed consent (as outlined in PLOS consent form) to publish this figure.

2.2 Participants

There were six participants. One participant was a female and one participant was one of the authors. Participants had normal hearing and vision. Participant age ranged from 21 to 33.

2.3 Stimuli

There were three blocks of experiments, each corresponding either to the audiovisual, the visual, or auditory experimental conditions. All subjects performed the audiovisual block first. The auditory and visual conditions were tested in one single session two months later. The visual and auditory blocks were counterbalanced in order across participants. The rationale behind this method was that all participants did the audiovisual localization trials—the main condition under study—without any prior training or knowledge of the stimuli. The two-month period between the sessions intended to create a gap in which the subjects would forget any learning occurred in the first session. Therefore, all conditions were run without interference of knowledge from other conditions.

Visual and auditory stimuli consisted of immersive reproductions of a young man playing an electronic portable organ in a room. In the visual condition only video, and no sound, was presented. In the auditory condition only sound was presented. In the audiovisual condition both video and sound were presented. Both visual and auditory stimuli were recorded in a large classroom, 10.9 m long by 6.5 m wide, and 2.2 m high. The reverberation time of the room was 1.9 sec. Stimuli were recorded along the 12.7 m room diagonal. Both the cameras and the microphone were positioned in one corner of the room, 1 m away from each wall, and at a height of 1.6 m. Recordings were taken of a male playing the C chord on an organ at different positions along the room diagonal.

The visual stimuli were recorded with 6 GoPro cameras mounted on a cubical support Freedom 360. With Autopano Video software the 6 videos were stitched together and thus a spherical 360 deg video was obtained. Visual stimuli were reproduced with an Oculus Rift device and allowed for free head movement with realtime rendering of the room. Auditory stimuli were recorded with an Eigenmike microphone (32 channels) reduced to first-order B-format (4 channels) rotating the axis in order to match video and audio directions. The Eigenmike microphone was chosen due to its structure, which has higher aliasing frequency than other B-format microphones available. Recording was rendered to create a 3D real-time audio environment using Directional Audio Coding [22, 23]. The auditory stimuli were reproduced through a set of Sennheiser HD 650 headphones. The audio was synchronized with the video using a cross correlation function. Recordings lasted for 3.3 sec and the organ was played continuously for 2 sec, starting at sec 0.7. The reason for the long stimulus duration had to do with the need to include enough information for subjects to access the direct-to-reverberant energy ratio and to access the full room reverberation. Both cues are known to be critical in auditory distance perception. There was a metronome playing in the background, which was positioned at the same distance as the organ. Its purpose was to time the keypress in all recordings. The visual stimuli consisted of random presentations of the spherical recording of the organ being played at 1, 3, 5, 7 and 9 m from the camera (Fig 2). The auditory stimuli consisted of random presentations of the sound environment of the organ being played at every meter, from 1 to 10 m in distance from the microphone. The audiovisual stimuli consisted of all the possible combinations of visual and auditory stimuli, randomly presented. In all conditions there were six repetitions per stimulus.

2.4 Apparatus and Procedure

Experiments took place in an acoustically treated room. Participants were seated on a chair, with the Oculus Rift placed over the eyes and the headphones over the Oculus Rift (Fig 2). The image was centered so that the stimuli were presented straight ahead. The visual, auditory, and audiovisual conditions were tested in separate sessions. The audiovisual condition was tested in two sessions. Before the beginning of each condition participants had a practice block. In the practice block, each stimulus was presented once in random order. Participants were instructed to pay attention to the room and to the stimuli, and to provide responses in the response interface. The response interface consisted of an iPad containing two sliders and one Continue button. One slider allowed to input responses to the visual stimuli, and the other one to the auditory stimuli. There was only one slider in the visual and auditory condition. The iPad interface and input could be seen on the Oculus Rift. By moving the slider participants could choose any value, ranging from 0 to 10. Participants were told that 0 corresponded to their own position in space, that 10 corresponded to a position just before the corner at the other end of the room, and that the values corresponded to actual meters in the room. Participants were asked to always answer to both visual and auditory distance after each audiovisual

trial, but they were allowed to choose which one to respond to first. They were specifically asked to always provide the same answer if both stimuli were perceived as stemming from the same point in space. This was important, as the rate of matching answers was used to compute the probability of perceived common cause in the causal inference model. Participants were further asked to provide the most honest report of what they actually perceived, not what they believed was the correct answer. This instruction had to do with the fact that there was no time limit to provide answers. It aimed specifically at asking subjects to avoid developing theories of what the experimental design might be and to avoid trying to find the correct answer, instead focusing in simply reporting percepts. Each trial started after pressing the Continue button.

2.5 Statistical Tests

To analyze the effect of each sensory modality over the visual and auditory distance estimates, a Kruskal-Wallis test was used. The choice of a non-parametric test had to do with the small number of subjects (n = 6). To test the multisensory models, a simple linear regression test was used. All pooled data were used, organized by subject, corresponding to the response counts per each of the 10 points in space in each of the 50 stimulus pairs per subject, fit to the corresponding predicted response counts (10 distances * 50 stimuli * 6 subjects = 3000 data points). Residuals were found to be normally distributed. Linear regressions were used because data were linearly related to the models. No model correction, such as Akaike's criterion, was used because all models had the same number of parameters in the fitting procedure and no generative model was used. Separate linear regressions were calculated for each model, for the visual estimates/predictions, the auditory estimates/predictions, and both. Separate regressions were also calculated for each subject.

3 Results

Overall, the localization of visual stimuli in distance was more accurate than that of the auditory stimuli. In the unimodal visual condition the average localization error was 0.34 m (0.49 SD), while in the unimodal auditory trials, the average localization error was 1.42 m (1.28 SD). In audiovisual trials, visual distance estimates had an average error of 0.33 m (0.51 SD). The auditory distance estimates in that condition had an average error of 1.48 m (1.33 SD). In bimodal trials, the perceived auditory distance was significantly affected both by auditory stimulus distance ($\chi^2_{(9,5)} = 1040.01$, $p = 0.000$) and visual stimulus distance ($\chi^2_{(9,5)} = 20.52$, $p = 0.000$). However, the perceived visual distance was only affected by visual stimulus distance ($\chi^2_{(9,5)} = 1731.23$, $p = 0.000$) and not by the auditory stimulus ($\chi^2_{(9,5)} = 0.28$, $p = 1.000$).

Multisensory integration was modeled using *causal inference*. For each stimulus pair, we: 1) calculated the mean (Eq (1)) and variance (Eq (2)) of the percept for a single underlying cause; 2) calculated the mean (Eq (3)) and variance (Eq (4)) of the percept for two underlying causes; 3) quantified the probability of common and separate causes (Eq (5)); and 4) calculated the final percept accounting for all previous steps (Eqs (6)–(9)). In the calculation of the final percept three strategies were tested: probability matching (Eq (6)), model averaging (Eqs (7) and (8)) and model selection (Eq (9)). Three additional models were tested: *sensory dominance* (Eq (10)), *mandatory integration* (Eqs (1) and (2)) and *no interaction*.

A visualization of the multisensory mechanisms simulated in each model is presented in Fig 1. The predicted distributions for each model were generated and tested against the actual response distributions. In Fig 3 all of the obtained auditory distance responses $\hat{s}_A$ in each multisensory trial type are presented together with the predictions from the probability matching strategy. The visual and auditory unimodal pooled distributions are also presented at the

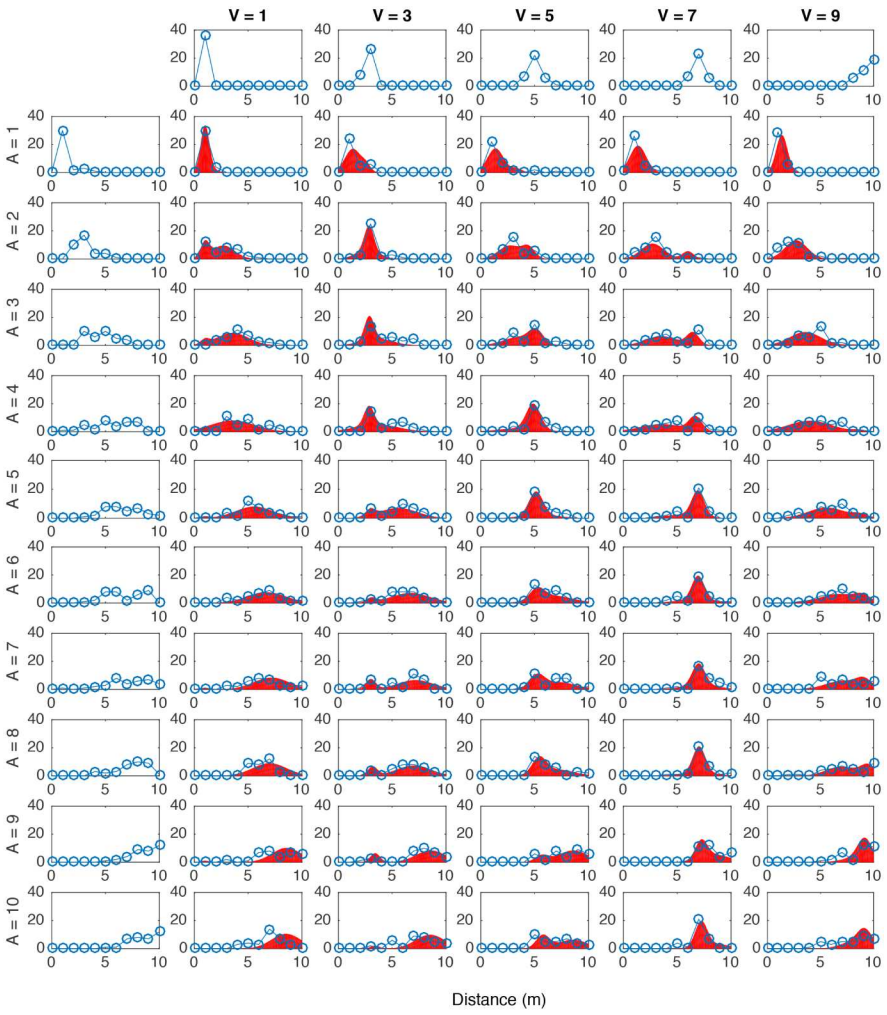

Fig 3. Auditory distance response distributions in all bimodal conditions (blue connected circles) and response distributions as predicted by *Causal Inference* resolved with the probability matching strategy (red area). The topmost graphs correspond to the unimodal visual distance distributions. The leftmost graphs correspond to the unimodal auditory distance distributions. Distributions were obtained by pooling all responses from all subjects. Auditory stimulus distance ranges in rows from 1 m distance (A = 1) to 10 m (A = 10) and visual stimulus distance ranging in columns from 1m to 9 m (V1 to V9 respectively).

doi:10.1371/journal.pone.0165391.g003

topmost and leftmost graphs, respectively. It is observed that the probability matching strategy constitutes a close approximation to the average auditory distance responses.

Figs 4 and 5 present visual and auditory distance responses in a sample of trial types against the predictions from the probability matching, model averaging and model selection strategies. In Table 1 a summary of all model fitting is presented. Looking at all data together (All), the *causal inference* models explained the data better, followed very closely by the *no interaction model*. The probability matching strategy yielded the best fits. This was also true in the data from each individual subject: for all subjects, the best fits were obtained with the *causal inference* model resolved with probability matching. In general, better fits were obtained for the visual distance estimates than for the auditory estimates. This may be related to the fact that visual estimates were more consistent and had lower variance that auditory estimates. The

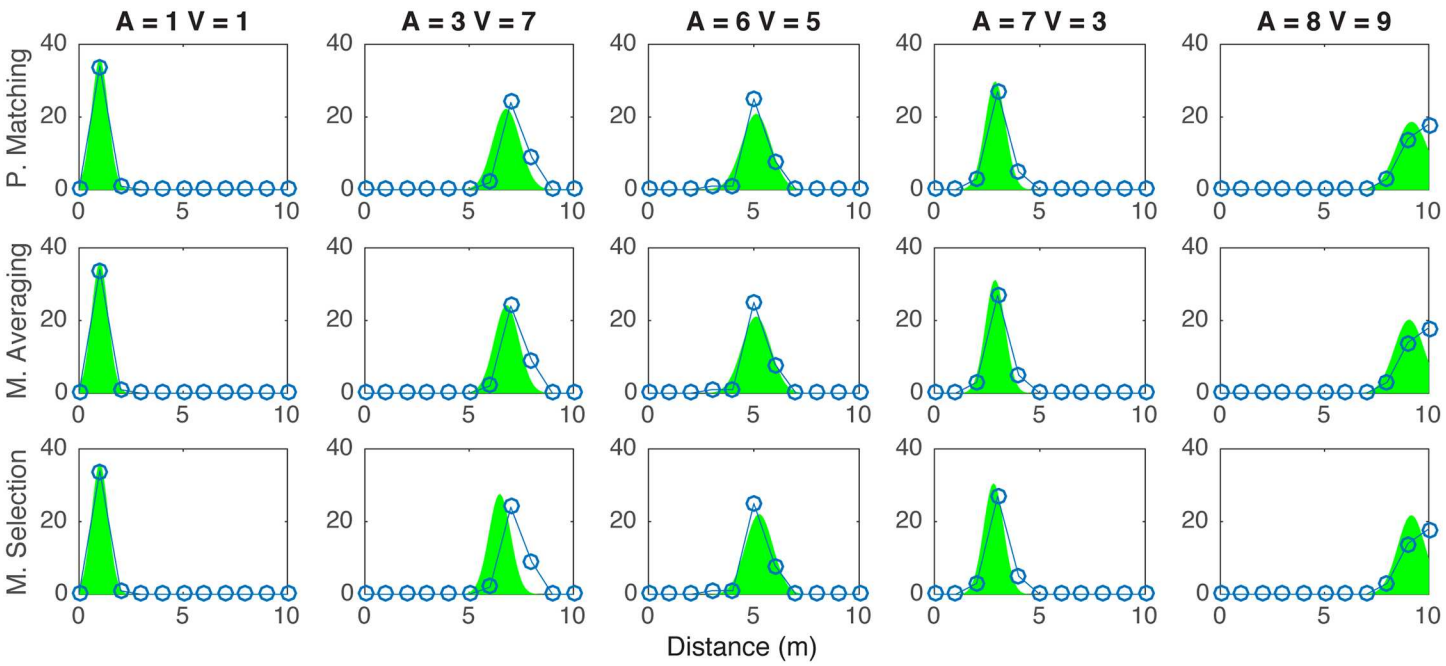

Fig 4. Five examples of visual distance response distributions (blue connected circles) and corresponding response distributions as predicted by *Causal Inference* resolved with the probability matching, model averaging, and model selection strategy (green area). Response distributions obtained from all pooled data.

doi:10.1371/journal.pone.0165391.g004

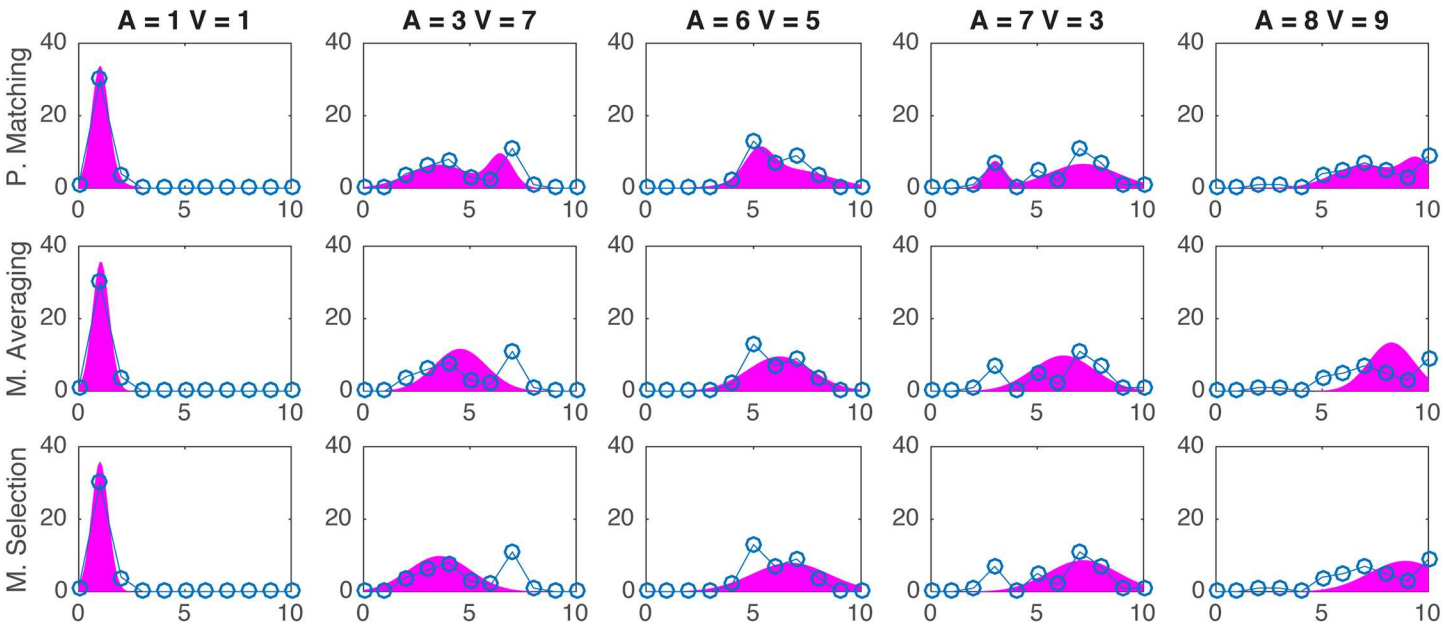

Fig 5. Five examples of auditory distance response distributions (blue connected circles) and corresponding response distributions as predicted by *Causal Inference* resolved with the probability matching, model averaging, and model selection strategy (magenta area). Response distributions obtained from all pooled data.

doi:10.1371/journal.pone.0165391.g005

Table 1. Goodness of fit of each model (r^2).

Model	All	All (A)	All (V)	s1	s2	s3	s4	s5	s6
SensoryDominance	0.451	0.121	0.853	0.435	0.462	0.66	0.530	0.411	0.447
MandatoryIntegration	0.448	0.074	0.952	0.378	0.475	0.469	0.506	0.410	0.472
NoInteraction	0.838	0.613	0.838	0.934	0.874	0.780	0.775	0.755	0.851
CausalInferencePM	**0.863**	**0.679**	0.963	**0.938**	**0.885**	**0.804**	**0.826**	**0.792**	**0.883**
CausalInferenceMA	0.841	0.618	**0.964**	0.935	0.869	0.778	0.790	0.743	0.872
CausalInferenceMS	0.849	0.642	0.963	0.935	0.872	0.787	0.795	0.790	0.867

Goodness of fit of each model against all distance estimates (All), the auditory distance estimates (All (A)) and the visual distance estimates (All (V)). Causal inference was tested with three strategies: probability matching (PM), model averaging (MA), and model selection (MS). The goodness of fit was obtained by simple linear regression with ordinary least squares regression. The regression analysis fit predictions against observed response rates per stimulus type and subject.

doi:10.1371/journal.pone.0165391.t001

visual estimates were well predicted by all tested models, and the best fit values was obtained with the *causal inference* model resolved with model averaging.

In a final step, sensory weights were calculated accounting for causal inference. The sensory weights were calculated as described in section 2B (Eqs (13) and (14)). They were calculated for each stimulus pair by averaging across all pooled data. The calculation followed the steps: 1) calculating each sensory weight for a single underlying cause; 2) calculating each sensory weight for two underlying causes; and 3) calculating the average of each weight accounting for the probability of each causal structure. In Fig 6 the sensory weights for the auditory distance estimates are presented for each stimuli pair. It can be observed that at all distances there is a window within which the visual cue largely affected the auditory estimate. This window of interaction is mobile and centered around the auditory stimulus position, meaning that when both stimuli were in close proximity they interacted in the formation of the auditory estimate. In those cases the auditory estimate was pulled in the direction of the visual stimulus. We can quantify the window size as the area where the weight of the visual cue surpasses that of the auditory cue. We find that the smallest window is observed with the auditory stimulus at 1 m, and it is 1 m wide. The largest interaction window occurs when the auditory stimulus is at 5 m, and is 4 m wide. The average interaction window is approximately 3 m, and stimuli interact little outside of it. The sensory weights in the visual distance estimates were markedly different. At all distances, the visual cue was the most weighted. The highest visual weights were found when the visual stimulus was at 1 m, where they were always equal to 1. The lowest weights were observed with the visual stimulus at 9 m, where the visual weight was close to 0.8 when the auditory stimulus was at 9 and 10 m.

4 Discussion

4.1 Audiovisual Distance

We studied the perception of distance from visual and auditory stimulation. It is remarkable that so little attention has been paid to multisensory interactions taking place in the extrapersonal space [24]. Little is known about the cue interactions in the audiovisual perception of distance. It is known that for visual and auditory events to be perceived as synchronized the auditory stimulus must lag the visual stimulus accounting for sound propagation velocity [25–27]. Previous research shows some potential audiovisual interactions. A cueing sound at the same distance as a visual target enhanced the detection of the visual target [28]. It has also been

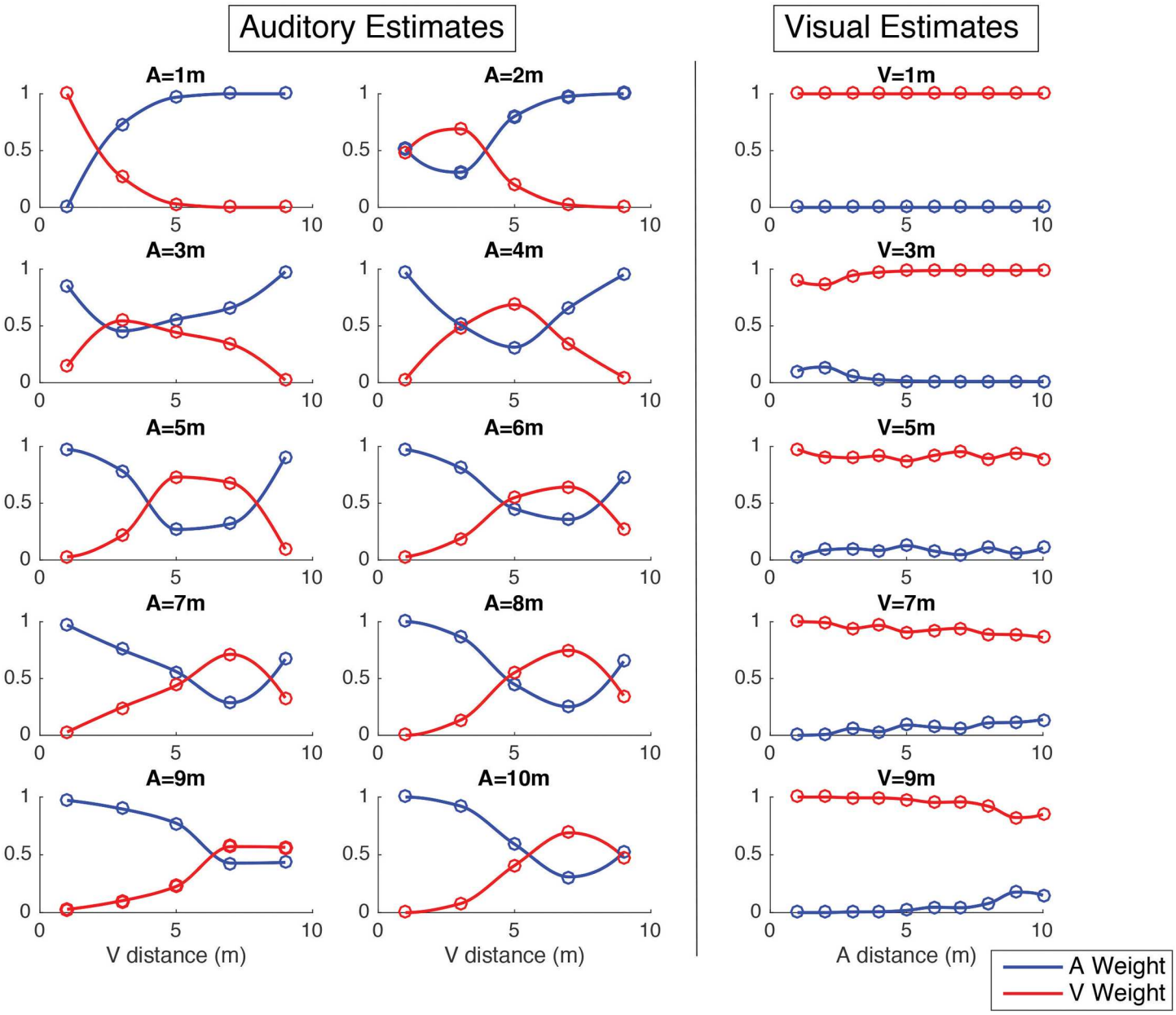

Fig 6. Sensory weights in auditory distance estimates. Calculated from all pooled data.

doi:10.1371/journal.pone.0165391.g006

reported that having visual references about the space improves auditory distance localization [29]. Seeing all response options, namely seeing an array of loudspeakers, improves localization of sounds in distance [30], while seeing only one loudspeaker biases the perceived distance toward it [31–34]. In those studies there were never visual events to relate the auditory events to. Therefore, the localization of visual and auditory events in distance when presented together had never been analyzed. The impact of sound events on visual distance perception remained unexplored, too, and it was not known how visual and auditory cues interact under congruent and incongruent conditions, and at several distances. In our experiment, subjects

were immersed in an audiovisual recorded room and were allowed to move their heads, since stimuli were generated in realtime. Therefore all sensory distance cues were available, except for binocular rivalry.

We found that visual distance estimates were for the most part accurate and that the impact of auditory cues was negligible. Auditory information had a maximum weight of 20 percent over the visual distance estimate, namely when stimuli were presented in proximity of each other and at large distances. In the bimodal trials it was found that vision has a large impact over the perceived auditory distance, contributing with more that 50 percent of the weight, namely when stimuli are presented within 3 m of each other. With larger stimuli separations the perceived auditory distance is mostly unaffected by visual stimulation. In light of the *causal inference* model, this window of interaction can be interpreted as the window within which humans alternate between inferring one and two separate external causes.

The fact that auditory cues had such low impact on visual distance estimates, and that visual cues only affected the auditory estimates within well defined windows may be related to the choice of stimuli employed in this experiment. In fact, stimuli were relatively longer in duration when compared to other audiovisual localization experiments. This choice of duration had to do with the fact that two main cues for auditory distance localization are reverberation time and direct-to-reverberant energy ratio [35–39], and reverberation as cue has a longer duration. Other auditory distance cues include sound level, auditory parallax, high frequency components and high room reflectance [35, 37, 40–43]. Our study included all of the above cues. This may certainly have increased the accuracy of the auditory image and therefore reduced uncertainty, which in turn may have contributed to lowered cue integration, higher rates of perceived separate causes, and generally low levels of cue interaction. In a similar way, our visual stimuli were very rich in distance cues: retinal size, familiar size, parallax, optic flow, texture gradient, and light and shade were available (see [44] for an overview on visual distance cues) and stimuli duration was enough for all cues to be processed with accuracy. Therefore, perhaps if stimuli duration was shorter, or if stimuli were impoverished and abstract, higher perceived common causality could have been obtained, or higher interaction observed. On the other hand, the use of realistic stimuli under well-controlled conditions can provide an insight on the mechanisms of multisensory combination as they often occur. It also provides for a preservation of stimulus identity from a common audiovisual source, which promotes multisensory binding [45]. From this point of view it is equally arguable that realistic and congruent stimuli could be associated to higher levels of perceived common causality and integration than if more abstract stimuli were used.

4.2 Multisensory Models

To assess the multisensory interactions in distance perception we tested several multisensory models. One of these models, the *causal inference* model, was tested with a different approach from previous multisensory studies. The model predicted that when separate causes are inferred the estimates have the same distribution as in the unisensory condition. When common causes are inferred it predicted that the distribution of the estimates can be given by the MLE model. It also proposed possible distributions of the combined estimates. While other distributions may be equally or more valid, we found that all tested *causal inference* models fit the data well. Three other longstanding models that do not assume causal inference were tested. It was found that overall *causal inference* fit better to the data, closely followed by the *no interaction* model. *Sensory dominance*, the longest standing model, explained the smallest portion of the overall data. *Mandatory integration*, which is still widely used as the main current model in multisensory processing, was the second worst. Note that mandatory integration is

also a part of the causal inference model. In the *causal inference* model it would be expected to occur if a common source was often inferred. Therefore, it is likely that the poor performance of the *mandatory integration* model in explaining the results from our experiment has to do with the low rates of perceived common causality. The *no interaction* model explained more than 80% of the overall data and did considerably well across all subjects. This model, too, is partly integrated in the *causal inference* model. It is expected to take place when two causes are inferred. Since in the experiment we report there were high rates of perceived separate causes, the good performance of the *no interaction* model is not surprising. The *causal inference* yielded the best fits in every tested case. Resolved with the probability matching strategy, it was found to explain the largest proportion of the overall data, which is in line with the finding of the study that proposed these resolving strategies [20]. In fact, when observing the response distributions of the auditory distance estimates it is often observed that there are two peaks. This bimodal distribution corresponds well to the estimates of each causal structure and it is a key feature in the *probability matching* strategy. It can be therefore hypothesized that, instead of combining each of the estimates from the inferred causal structures, perceivers alternate their response strategy between one and the other. All subjects seemed to follow more the probability matching strategy, which is not to say they did not alternate between strategies during the experiment. Indeed, looking at the visual distance estimates separately, we see that model averaging did slightly better, although all models fit very similarly. [20] compared the three causal inference strategies in a large population of subjects for localization of audiovisual horizontal stimuli. They found that the majority of the subjects followed more the probability matching strategy, while others followed more a model averaging or model selection strategy. A model averaging strategy means that subjects respond mostly according to a linear weighted average on the two causal structures, while a model selection strategy means that subjects respond mostly according to the most probable causal structure. It must be noted that here these strategies were tested with different calculations from the study by [20]: no generative model was used, and instead model predictions were obtained by using the same calculations for response centroid, but original calculations for response distribution. Therefore result comparisons between studies should be read with caution. The models used here seem however to work as a plausible alternative to test the *causal inference* model with much reduced computational complexity and demand.

In a final step, with the purpose of quantifying the overall importance of each sensory cue over the other, we proposed a new method to calculate sensory weights. This computation allows for an intuitive visualization of the interactions between cues in all tested cue combination cases. It also allows for the quantification of the relevance of each cue, and for the measurement of the window of interaction. Here, the window of interaction was defined as the stimuli range within which the sensory estimates of one sensory modality were affected in more than 50% by another concurrent sensory modality. In this case, it was observed that there is a clear window of interaction in auditory distance perception, which is always centered around the stimulus position itself. However, there is no such window in the perception of visual distance.

Taking an overview of our data, they seem to suggest that visual distance and auditory distance percepts are formulated though different mechanisms. The sensory weights of each cue are different for visual and for auditory distance estimates, even when they are mostly perceived as co-localized. Also, the multisensory model that best explains estimates in one sensory modality might not be best suited for the other sensory modality. This possibility calls for the need of analysing the multisensory percepts in each modality separately, and for the testing of multisensory models with this in mind.

Several notes should be taken while reading our test of multisensory models in the perception of audiovisual distance. Firstly, it must be noted that the performance of each tested model against our data may have been influenced by the research method itself. It may be that the fact that only 10% of the trials presented co-localized stimuli decreased the probability of inferred common causality. It may also be that the availability of very clear distance cues combined with a long stimulus duration might have decreased the level of cue interaction. However, it is also true that currently all research testing the *causal inference* model uses similarly low rates of congruent stimuli. Indeed, even articles testing other multisensory mechanisms such as the MLE tend to exhibit this limitation. In any case, it remains an open question whether the rate of congruent stimuli would affect model performance, and future studies should address this issue. It may also be that the availability of most distance cues, combined with long stimulus duration, may have affected the level of cue interaction. This is also an open question in most research experiments in this field, and only a new battery of tests manipulating stimulus quality would be able to answer it. Nevertheless, the *causal inference* model is sensitive to changes in rate of cue integration, by simply assuming different values of probability of common cause, and it can therefore be expected that the model would perform equally as well with different stimuli arrangements and experimental designs.

Finally, the proposed *causal inference* model was merely a first suggestion of how perceptual causal inference mechanisms might be estimated with simple mathematical formulations. There is the need to further look into this model and explore the existence of potential biases, test alternative distributions, and alternative resolution strategies. The proposed model should also be tested against other *causal inference* model formulations and other datasets.

5 Conclusion

The models of multisensory integration are a very useful tool to describe quantitatively how different sensory cues interact to produce a sensory estimate. Here, these models are brought forward to explain the data from an experiment on the audiovisual perception of distance. The *causal inference* model with probability matching strategy approximated the overall data better than the other models. The causal inference principles can also be used to calculate the sensory weights under all stimuli combinations. These weights revealed that, within a given window, the visual cue has greater importance than the auditory cue in the perception of auditory distance. The visual cue is the most prominent one in the perception of visual distance.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska Curie grant No 659114 and by the Academy of Finland, decision no 266239.

Author Contributions

Conceptualization: CM.

Data curation: CM PM.

Formal analysis: CM PM.

Funding acquisition: CM.

Investigation: PM CM.

Methodology: CM.

Project administration: CM.

Resources: VP CM.

Software: PM.

Supervision: CM VP.

Validation: CM.

Visualization: CM PM.

Writing – original draft: CM.

Writing – review & editing: CM.

References

1. Welch Robert B and Warren David H. Immediate perceptual response to intersensory discrepancy. Psychological bulletin, 88(3):638, 1980. doi: 10.1037/0033-2909.88.3.638 PMID: 7003641
2. Colavita Francis B. Human sensory dominance. Perception & Psychophysics, 16(2):409–412, 1974. doi: 10.3758/BF03203962
3. Morein-Zamir Sharon, Soto-Faraco Salvador, and Kingstone Alan. Auditory capture of vision: examining temporal ventriloquism. Cognitive Brain Research, 17(1):154–163, 2003. doi: 10.1016/S0926-6410(03)00089-2 PMID: 12763201
4. Vroomen Jean and de Gelder Beatrice. Temporal ventriloquism: sound modulates the flash-lag effect. Journal of Experimental Psychology: Human Perception and Performance, 30(3):513, 2004. doi: 10.1037/0096-1523.30.3.513 PMID: 15161383
5. Pavani Francesco, Spence Charles, and Driver Jon. Visual capture of touch: Out-of-the-body experiences with rubber gloves. Psychological science, 11(5):353–359, 2000. doi: 10.1111/1467-9280.00270 PMID: 11228904
6. Choe Chong S, Welch Robert B, Gilford Robb M, and Juola James F. The ventriloquist effect: Visual dominance or response bias? Perception & Psychophysics, 18(1):55–60, 1975. doi: 10.3758/BF03199367
7. Ernst Marc O and Banks Martin S. Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 415(6870):429–433, 2002. doi: 10.1038/415429a PMID: 11807554
8. Alais David and Burr David. The ventriloquist effect results from near-optimal bimodal integration. Current biology, 14(3):257–262, 2004. doi: 10.1016/j.cub.2004.01.029 PMID: 14761661
9. Battaglia Peter W, Jacobs Robert A, and Aslin Richard N. Bayesian integration of visual and auditory signals for spatial localization. Journal of the Optical Society of America A, 20(7):1391–1397, 2003. doi: 10.1364/JOSAA.20.001391
10. Deneve Sophie and Pouget Alexandre. Bayesian multisensory integration and cross-modal spatial links. Journal of Physiology-Paris, 98(1):249–258, 2004. doi: 10.1016/j.jphysparis.2004.03.011 PMID: 15477036
11. Andersen Tobias S, Tiippana Kaisa, and Sams Mikko. Maximum likelihood integration of rapid flashes and beeps. Neuroscience Letters, 380(1):155–160, 2005. doi: 10.1016/j.neulet.2005.01.030 PMID: 15854769
12. Shams Ladan, Ma Wei Ji, and Beierholm Ulrik. Sound-induced flash illusion as an optimal percept. Neuroreport, 16(17):1923–1927, 2005. doi: 10.1097/01.wnr.0000187634.68504.bb PMID: 16272880
13. Elliott Mark T, Wing AM, and Welchman AE. Multisensory cues improve sensorimotor synchronisation. European Journal of Neuroscience, 31(10):1828–1835, 2010. doi: 10.1111/j.1460-9568.2010.07205.x PMID: 20584187
14. Mendonça Catarina, Santos Jorge A, and López-Moliner Joan. The benefit of multisensory integration with biological motion signals. Experimental brain research, 213(2-3):185–192, 2011. doi: 10.1007/s00221-011-2620-4 PMID: 21424256
15. Roach Neil W, Heron James, and McGraw Paul V. Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. Proceedings of the Royal Society of London B: Biological Sciences, 273(1598):2159–2168, 2006. doi: 10.1098/rspb.2006.3578 PMID: 16901835

16. Sato Yoshiyuki, Toyoizumi Taro, and Aihara Kazuyuki. Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli. Neural computation, 19(12):3335–3355, 2007. doi: 10.1162/neco.2007.19.12.3335 PMID: 17970656
17. Beierholm Ulrik, Shams Ladan, Ma Wei J, and Koerding Konrad. Comparing bayesian models for multisensory cue combination without mandatory integration. In Advances in neural information processing systems, pages 81–88, 2007.
18. Shams Ladan and Beierholm Ulrik R. Causal inference in perception. Trends in cognitive sciences, 14 (9):425–432, 2010. doi: 10.1016/j.tics.2010.07.001 PMID: 20705502
19. Körding Konrad P, Beierholm Ulrik, Ma Wei Ji, Quartz Steven, Tenenbaum Joshua B, and Shams Ladan. Causal inference in multisensory perception. PLoS ONE, 2:e943, 09 2007. doi: 10.1371/journal.pone.0000943 PMID: 17895984
20. Wozny David R, Beierholm Ulrik R, and Shams Ladan. Probability matching as a computational strategy used in perception. PLoS Comput Biol, 6(8):e1000871, 2010. doi: 10.1371/journal.pcbi.1000871 PMID: 20700493
21. Odegaard Brian, Wozny David R, and Shams Ladan. Biases in visual, auditory, and audiovisual perception of space. PLoS Comput Biol, 11(12):e1004649, 2015. doi: 10.1371/journal.pcbi.1004649 PMID: 26646312
22. Pulkki Ville. Spatial sound reproduction with directional audio coding. Journal of the Audio Engineering Society, 55(6):503–516, 2007.
23. Mikko-Ville Laitinen and Ville Pulkki. Binaural reproduction for directional audio coding. In Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on, pages 337–340. IEEE, 2009.
24. Van der Stoep N, Nijboer TCW, Van der Stigchel S, and Spence C. Multisensory interactions in the depth plane in front and rear space: a review. Neuropsychologia, 70:335–349, 2015. doi: 10.1016/j.neuropsychologia.2014.12.007 PMID: 25498407
25. Silva Carlos César, Mendonça Catarina, Mouta Sandra, Silva Rosa, Campos José Creissac, and Santos Jorge. Depth cues and perceived audiovisual synchrony of biological motion. PloS one, 8(11): e80096, 2013. doi: 10.1371/journal.pone.0080096 PMID: 24244617
26. Alais David and Carlile Simon. Synchronizing to real events: Subjective audiovisual alignment scales with perceived auditory depth and speed of sound. Proceedings of the National Academy of Sciences of the United States of America, 102(6):2244–2247, 2005. doi: 10.1073/pnas.0407034102 PMID: 15668388
27. Sugita Yoichi and Suzuki Yôiti. Audiovisual perception: Implicit estimation of sound-arrival time. Nature, 421(6926):911–911, 2003. doi: 10.1038/421911a PMID: 12606990
28. Van der Stoep N, Van der Stigchel S, Nijboer TCW, and Van der Smagt MJ. Audiovisual integration in near and far space: effects of changes in distance and stimulus effectiveness. Experimental brain research, pages 1–14, 2015.
29. Calcagno E, Abregú Ezequiel, Eguía Manuel C, and Vergara RO. The role of vision in auditory distance perception. Perception, 41(2):175–192, 2012. doi: 10.1068/p7153 PMID: 22670346
30. Zahorik Pavel. Estimating sound source distance with and without vision. Optometry & Vision Science, 78(5):270–275, 2001. doi: 10.1097/00006324-200105000-00009 PMID: 11384003
31. Gardner Mark B. Proximity image effect in sound localization. The Journal of the Acoustical Society of America, 43(1):163–163, 1968. doi: 10.1121/1.1910747 PMID: 5636394
32. Gardner Mark B. Distance estimation of 0 or apparent 0 oriented speech signals in anechoic space. The Journal of the Acoustical Society of America, 45(1):47–53, 1969. doi: 10.1121/1.1911372 PMID: 5797146
33. Mershon Donald H, Desaulniers Douglas H, Amerson Thomas L, and Kiefer Stephan A. Visual capture in auditory distance perception: Proximity image effect reconsidered. Journal of Auditory Research, 1980. PMID: 7345059
34. Anderson Paul Wallace and Zahorik Pavel. Auditory/visual distance estimation: accuracy and variability. Frontiers in Psychology, 5 (1097), 2014. doi: 10.3389/fpsyg.2014.01097
35. Mershon DonaldH. and King L.Edward. Intensity and reverberation as factors in the auditory perception of egocentric distance. Perception Psychophysics, 18(6):409–415, 1975. doi: 10.3758/BF03204113
36. Zahorik Pavel. Assessing auditory distance perception using virtual acoustics. The Journal of the Acoustical Society of America, 111(4):1832–1846, 2002. doi: 10.1121/1.1458027 PMID: 12002867
37. Zahorik Pavel, Brungart Douglas S, and Bronkhorst Adelbert W. Auditory distance perception in humans: A summary of past and present research. Acta Acustica united with Acustica, 91(3):409–420, 2005.

38. Bronkhorst Adelbert W and Houtgast Tammo. Auditory distance perception in rooms. Nature, 397 (6719):517–520, 1999. doi: 10.1038/17374 PMID: 10028966

39. Moore David R and King Andrew J. Auditory perception: The near and far of sound localization. Current Biology, 9(10):R361—R363, 1999. doi: 10.1016/S0960-9822(99)80227-9 PMID: 10339417

40. Coleman Paul D. An analysis of cues to auditory depth perception in free space. Psychological Bulletin, 60(3):302, 1963. doi: 10.1037/h0045716 PMID: 14022252

41. Little Alex D, Mershon Donald H, and Cox Patrick H. Spectral content as a cue to perceived auditory distance. Perception, 21(3):405–416, 1992. doi: 10.1068/p210405 PMID: 1437460

42. Kim Hae-Young, Suzuki Yoiti, Takane Shouichi, and Sone Toshio. Control of auditory distance perception based on the auditory parallax model. Applied Acoustics, 62(3):245–270, 2001. doi: 10.1016/S0003-682X(00)00023-2

43. Mershon Donald H, Ballenger William L, Little Alex D, McMurtry Patrick L, and Buchanan Judith L. Effects of room reflectance and background noise on perceived auditory distance. Perception, 18 (3):403–416, 1989. doi: 10.1068/p180403 PMID: 2798023

44. Gilinsky Alberta S. Perceived size and distance in visual space. Psychological Review, 58(6):460, 1951. doi: 10.1037/h0061505 PMID: 14900306

45. Bizley Jennifer K, Maddox Ross K, and Lee Adrian KC. Defining auditory-visual objects: Behavioral tests and physiological mechanisms. Trends in neurosciences, 39(2):74–85, 2016. doi: 10.1016/j.tins.2015.12.007 PMID: 26775728