Integration of visual and spoken cues in a virtual reality navigation task

Serena DeStefani (serena.destefani@rutgers.edu)

Karin Stromswold (karin@ruccs.rutgers.edu)

Jacob Feldman (jacob@ruccs.rutgers.edu)
Dept. of Psychology, Center for Cognitive Science
Rutgers University - New Brunswick
Piscataway, NJ 08854 USA

Abstract

When integrating information in real time from multiple modalities or sources, such as when navigating with the help of GPS voice instructions along with a visual map, a decision-maker is faced with a difficult cue integration problem. The two sources, in this case visual and spoken, have potentially very different interpretations or presumed reliability. When making decisions in real time, how do we combine cues coming from visual and linguistic evidence sources? In a sequence of three studies we asked participants to navigate through a set of virtual mazes using a head-mounted virtual reality display. Each maze consisted of a series of T intersections, at each of which the subject was presented with a visual cue and a spoken cue, each separately indicating which direction to continue through the maze. However the two cues did not always agree, forcing the subject to make a decision about which cue to "trust." Each type of cue had a certain level of reliability (probability of providing correct guidance), independent from the other cue. Subjects learned over the course of trials how much to follow each cue, but we found that they generally trusted spoken cues more than visual ones, notwithstanding the objectively matched reliability levels. Finally, we show how subjects' tendency to favor the spoken cue can be modeled as a Bayesian prior favoring trusting such sources more than visual ones.

Keywords: multimodal integration; navigation; Bayesian inference; virtual reality

Navigation as a Decision-making Problem

Navigation, the process by which an agent chooses a path through the environment, has been studied from the point of view of cognitive strategy (Loomis, Klatzky, Golledge, & Philbeck, 1999; Golledge, 2003), computation (Durrant-Whyte & Bailey, 2006) and neural representation (Andersen, Snyder, Bradley, & Xing, 1997). The cognitive component of navigation—often referred to as *wayfinding*— requires the agent to integrate information from multiple sources (Deneve & Pouget, 2004), including both environmental cues and internal representations, in order to choose a path that optimizes success in arriving at a target location. In this work we treat navigation as a decision problem, focusing on how disparate and potentially conflicting information is integrated to arrive at a decision about which way to go.

Consider for example the problem faced by a driver navigating with the assistance of a GPS device (Holland, Morse, & Gedenryd, 2002). The driver must integrate visual cues about which way to turn (e.g. road structure and signage) with spoken instructions from the GPS explicitly indicating which way to turn at each intersection. If these cues conflict—for example, the GPS says to turn left when a visible sign suggests turning right—the driver must make a decision in the face of what is traditionally called *cue conflict*. In principle, in such a situation, the driver may have a prior beliefs about which cue is more reliable, which can be thought of as a Bayesian prior over the probability that each cue is likely to be accurate. Which cue will they "trust" more, and how do they integrate evidence from sources of various degrees of reliability? The experiment below is designed to answer these questions, with the goal of quantifying subjects' tendency to rely on each type of cue source in a Bayesian framework.

There is a substantial literature on cue combination (Landy, Maloney, Johnston, & Young, 1995), much of which suggests that human observers integrate cues in an approximately optimal fashion (Cheng, Shettleworth, Huttenlocher, & Rieser, 2007). In an optimal framework, each cue is weighted according to its reliability (for Gaussian cues, often quantified as the inverse of the variance, see Yuille & Bülthoff, 1996). Visual cues have often been regarded as dominant over other cues (Colavita, 1974; Egeth & Sager, 1977), though in more modern treatments this has been attributed to their greater reliability and thus higher weight (Ernst & Banks, 2002). In some contexts auditory cues have been found to predominate over visual ones (Grahn, Henry, & McAuley, 2011), though again this has been argued to reflect their greater reliability in those contexts (Burr, Banks, & Morrone, 2009). Hence while individual cues vary in reliability, human decisions seem generally to integrate them in an approximately rational fashion.

However the juxtaposition of visual and *spoken* (auditory linguistic) cues faced by the driver using the GPS—and fairly ubiquitous in other natural contexts—does not seem to have been addressed in the literature, and raises a number of novel issues. As discussed below, spoken cues may carry a subjective weight potentially disproportional to their objective reliability. So in the experiments we conducted, we asked subjects to navigate using a combination of visual and spoken cues, whose statistical reliability we controlled and manipulated. The overall goal was to understand how such cues are integrated, and in particular to quantify the degree of "trust" subjects allocate to different kinds of information sources.

Experiments

We created a set of nine mazes in a virtual reality (VR) environment which subjects experienced through a head-mounted display (Fig. 1). Each maze consists of a sequence of 32 T-junctions (see partial example, Fig. 1a) where the subject must choose to turn left or right, with one direction leading to a continuation of the maze (the correct choice) and the other to a dead end (the incorrect one).

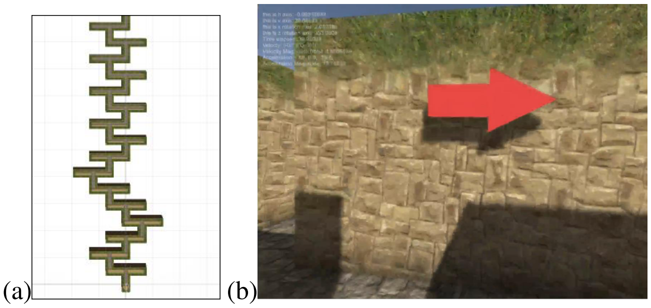


Figure 1: (a) Example overview of part of a sample maze, showing the sequence of T-junctions (b) Inside the virtual maze, showing visual cue (arrow) at a T-junction.

Subjects advanced through the maze using a keyboard, and turned left or right by turning their head. As the subject approached each intersection, they were presented with a visual cue indicating which direction to turn (e.g. in Exp. 1, a large red left or right arrow floating against the rear wall, Fig. 1b) and an spoken cue instructing them to "go left" or "go right" which was synchronized with the onset of the visual cue subjects listened to the spoken cue through headphones and the sound volume was the same on each side). Critically, the two cues were not 100% reliable (see details below) and not always in agreement with each other. Subjects were instructed to make a decision rapidly and to avoid attempting to look ahead before committing to a choice. After making a turn, the subject received feedback by seeing either an open corridor (indicating a correct choice) or a brick wall (an incorrect one). The main dependent variable was their sequence of choices (left or right) at each point in the maze, reflecting both their a priori trust in each cue as well as their judgment about the trustworthiness of each cue based on the accumulating evidence from previous trials. The choice was recorded manually by the experimenter.

We ran three versions of this experiment (Exps. 1–3) which differed in the way cues were presented, as explained below. In each of the experiments, each one of the nine mazes consisted of 32 T-intersections, at 16 of which the maze continued left and at 16 of which it continued right, in random order. (We kept the number of intersections low in order to limit trial length to about three minutes to minimize the risk of dizziness.) Within a given maze, each cue provided correct guidance on a fixed proportion of trials and incorrect guidance on the others, a proportion we refer to as the cue's *reliability*. (Note that we use this term to mean literally the probability of conveying the correct information, not as a synonym for the optimal cue weight as often done in connection to Gaussian cues.) Reliability levels were 25%, 50% and 75% for each cue type, with visual cue reliability crossed with spoken cue reliability across mazes for total of $3 \times 3 = 9$ reliability conditions. For example, on a given maze the visual cue might be 75% reliable but the spoken cue 25%, meaning that on $.75 \times .25 \times 32 = 6$ intersections both cues were correct, while on $.25 \times .25 \times 32 = 2$ intersections the visual cue was incorrect but the spoken one was correct, and so forth. Subjects were encouraged to consider each maze a completely new situation with cues to be evaluated *de novo*. To encourage participants to treat each maze as a completely new situation, the background colors were different in each condition. Each subject ran all nine conditions in random order, so across the entire experiment visual and spoken reliability rate were equated. Participants were undergraduate and graduate students at Rutgers University, all with normal hearing and normal or corrected-to-normal vision. Participants using glasses were excluded because of incompatibility with the VR headset. Each experiment was run on a completely new set of participants. Note that because of the nature of the task, which involves frequent accelerated movements and head turns, a sizable portion of the participants experienced dizziness and were unable to complete the task. This difficulty is partly responsible for the small number of subjects, in light of which we treat our conclusions as preliminary.

Experiment 1. In Exp. 1 the visual cue was a large red arrow (pointing left or right) projected against the back wall of the T-junction (Fig. 1b), and the spoken cue was a voice saying "go left" or "go right" recorded by a female native English speaker. In order to avoid giving either visual or spoken cues temporal priority, we used auditory software (Praat 6.0.16, Boersma & Weenink, 2018) to determine the point in time at which the voice initiated the key disambiguating word "right" or "left", and synchronized it with the onset of the visual cue.

Subjects, Exp. 1. Data were collected from 5 subjects for Exp. 1 (3 women, 2 men). Subjects were ignorant of the goal of the study and were paid for their participation. Because the use of the VR headset (Oculus Rift) can lead to nausea, participants were allowed to take breaks whenever they wanted. All the studies were approved by the Rutgers IRB.

Results, Exp. 1. We analyzed only discordant trials, i.e. trials on which visual and spoken cues disagreed. Fig. 2 shows plots of the proportion of visual-following vs spoken-following choices on such trials, as a function of cue reliability. The figure includes individual subject plots as well as an aggregate plot showing the mean over subjects. The results show a clear and consistent bias in favor of the spoken cue, illustrated by the offset in each plot of the red curve over the blue.

To test models comparing the effect of factors such as the nature of the cue and the cue's reliability level, we analyzed the number of times a participant chose a given cue in discordant trials with a 2 × 2 Bayesian ANOVA with default JZS priors and including a random effect for individual subjects (Rouder, Morey, Speckman, & Province, 2012). The factors were the type of cue (spoken or visual) and the reliability level

(25%, 50% or 75%).

Results showed that including either the cue ($\text{BF}_{10} = 18.22 \pm .14\%$) or the reliability level ($\text{BF}_{10} = 60.78 \pm .18\%$) as predictors improved the model with respect to the null model in which the variance is only explained by individual differences. A model including both cue and reliability as predictors was strongly favored over any other model ($\text{BF}_{10} = 45842.85 \pm .35\%$). The offset between spoken and visual cues was more clear-cut in some subjects (e.g. #1) than others (#4), but was present in each individual subject.

Subjects' compliance with each cue increased with its objective reliability, as would be expected. The more each cue gave the right answer, the more the subjects tended to follow it. However above and beyond this dependence on the actual statistical properties of each cue, the subjects exhibited a prior bias to find spoken instructions more believable than visual guidance. We investigate this bias more carefully in the modeling section below.

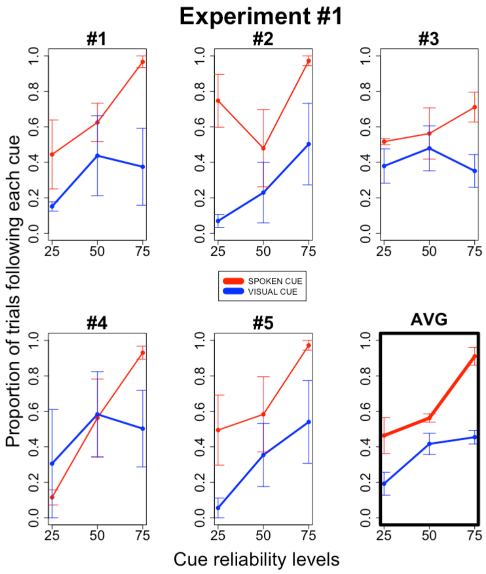


Figure 2: Plots showing the proportion of trials on which subjects followed the visual cue (blue) or the spoken cue (red) in the discordant cases in Exp. 1, across reliability levels (abscissa).

Experiment 2. It is possible that phonetic modulation of the word "go" in the phrases "go left" or "go right" could give subjects an early indication of the upcoming direction, thus giving the spoken cue temporal priority (Allopenna, Magnuson, & Tanenhaus, 1998). Hence in Exp. 2 we repeated Exp. 1 except without the word "go" in the spoken instruction. Instead spoken cues consisted simply of the word "left" or "right," recorded by a female native English speaker.

Subjects, Exp. 2. Data were collected from 6 participants in Exp. 2 (4 women, 2 men). Again, all participants were ignorant of the goal of the study and were paid for their participation.

Results, Exp. 2. Fig. 3 shows plots of the proportion of visual-following vs spoken-following choices as a function of cue type and cue reliability on discordant trials. As can be seen in the plots, both cue type and reliability show effects in the same direction as in Exp. 1, but both effects are now swamped by noise, and neither a model based on cue nor a model based on reliability is favored over the null model (cue type: $\text{BF}_{10} = .67 \pm .28\%$; reliability $\text{BF}_{10} = .56 \pm .19\%$). Evidently, the modification of the spoken instruction substantially diminished its value as a cue, perhaps because the one-word version ("left" or "right") is no longer a syntactically or pragmatically well-formed utterance.

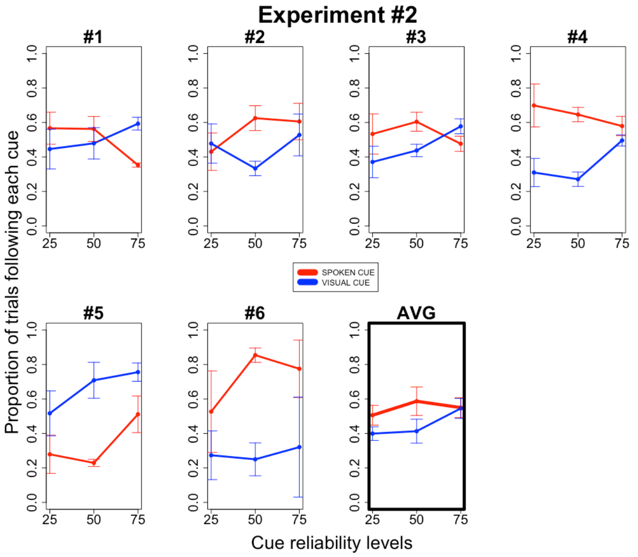


Figure 3: Plots showing the proportion of trials on which subjects followed the visual cue (blue) or the spoken cue (red) in the discordant cases in Exp. 2, across reliability levels (abscissa).

Experiment 3. We wondered whether the unexpectedly low weight our subjects placed on the visual cue was because the cue itself, a static arrow, was simply not salient enough to compete with spoken instructions. Motion is a simple manipulation that is known to increase visual salience. For example the onset of motion attracts attention in both adults (Abrams & Christ, 2003) and infants (Farroni, Johnson, Brockbank, & Simion, 2000). So in Exp. 3 we increased the salience of the visual cue by replacing the static arrow used in Exps. 1 and 2 with a moving arrow "flowing" in the indicated direction. The goal was simply to check whether this change might tip the balance of our subjects' decisions in favor of the visual cue. Exp. 3 was identical to Exp. 1 except for the nature of the visual cue, which was now an arrow sliding smoothly along the back wall of the T-junction in the indicated direction. Maze construction, subject instructions, reliability levels, and spoken cue presentation were all as in Exp. 1.

Subjects, Exp. 3. Data were collected from 7 participants in Exp. 3 (2 women, 5 men). Again all participants were ignorant of the goal of the study and were paid for their participation.

Results, Exp. 3. The results of Exp. 3 are shown in Fig. 4, again showing the proportion of visual-following vs spoken-following choices a function of cue type and cue reliability

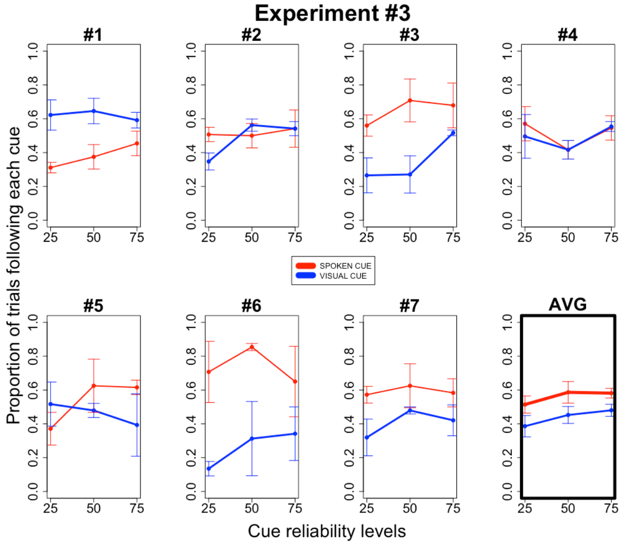


Figure 4: Plots showing the proportion of trials on which subjects followed the visual cue (blue) and the spoken cue (red) in the discordant cases in Exp. 3, across reliability levels (abscissa).

on discordant trials. As in Exps. 1 and 2, most subjects show a marked bias in favor of believing the spoken cue, on top of the effect of objective statistical properties of each cue. In the Bayesian ANOVA, the model including cue as an explanatory variable was favored moderately over the null model ($\mathrm{BF}_{10} = 6.8 \pm 2\%$), while the model including reliability was not ($\mathrm{BF}_{10} = .68 \pm 1\%$).

Hence these results again corroborate the basic finding of a bias in favor of the spoken cue over the visual one, although for some reason these subjects were unable to learn effectively from feedback. Indeed the bias in favor of the spoken cue seems to survive even when the salience of the visual cue is substantially enhanced.

Modeling subjects' bias

As our subjects progressed through each maze, they had to judge the reliability of each cue source. At the start of each maze, they had no evidence on which to base a judgment—only whatever general expectations they might have about source reliability. But as they progressed they accumulated evidence about the statistical reliability of each cue type. Our results above show that subjects are at least somewhat able to learn from this evidence, in that their overall reliance on each source increased monotonically with its objective statistical reliability: the more often a source was correct, the more it was trusted. However their integration of the objective evidence was apparently also tempered by some bias, in that they tended to trust spoken instructions more than visual ones even though their reliabilities were objectively the same. In this section we attempt to quantify this bias more precisely using a Bayesian modeling framework.

In our experimental framework, each cue is a binary source, on each trial randomly generating either a correct instruction ("success") or an incorrect one ("failure") (cf. Backus, 2009). The reliability of the source is simply the probability λ of success, i.e. of issuing correct instructions. If the subject assumes that successive trials are independent, as is actually correct in our experiment, then this is a simple example of a Bernoulli estimation problem, in which the goal is to estimate the success probability λ from a sequence of binary samples. That is, the subject observes a sequence of correct and incorrect instructions from a given source, and on that basis forms an estimate of the probability of correct instructions from that particular source going forward. In what follows we model this learning process for our subjects, in particular attempting to quantify the *bias* that they bring to bear on this estimation problems—that is, any tendency they might have infer high or low values of the reliability parameter.

In a Bayesian framework, the bias is expressed by the prior $p(\lambda)$ over possible reliability levels λ. This distribution expresses how plausible the observer finds each potential level of reliability before collecting evidence—for example placing more probability mass on higher values of λ, indicating a bias to trust the source, or on lower values, indicating a bias to distrust it. Hence below, for each subject on each maze, we quantify each subject's bias for each cue type by estimating the prior distribution $p(\lambda)$ most consistent with their sequence of decisions through the maze. This is the prior that best explains the choices they actually made.

In Bayesian treatments, the prior over the Bernoulli probability parameter is usually assumed to follow a Beta distribution $\mathrm{Beta}(\alpha,\beta)$, which is conjugate to the binomial likelihood distribution which follows from the sampling model (Lindley & Phillips, 1976; Griffiths, Kemp, & Tenenbaum, 2008). The parameters α and β jointly characterize the distribution, respectively modulating the bias for success (correct information) and failure (misinformation). More specifically α and β can be thought of as respectively the number of successes and failures observed prior to the first sample. For example a learner with a prior of Beta(7,2) has a bias equivalent to actually having seen 7 correct instructions and 2 incorrect instructions before the experiment began. In this sense the sum $\alpha+\beta$ represents the "confidence" the subject has in his or her prior (i.e. the total number of previous observations the prior is equivalent to) while the difference $\alpha-\beta$ represents the *direction* of the bias (i.e. the degree to which the prior favors trust over mistrust of the source). In (α,β) space (plotted in figures below), the region $\alpha > \beta$ corresponds to a bias in favor of trust, and the region $\alpha < \beta$ to bias to distrust it, and the diagonal $\alpha = \beta$ corresponds to perfect neutrality. By estimating where a particular subject on a particular maze falls in this space, we can precisely characterize the nature and direction of their attitudes towards each type of cue source.

Note that this method of quantifying bias is broadly applicable to any situation in which an observer must estimate the reliability of a binary cue source, a situation exemplified by our experiments but also very common in a range of real-life contexts. Note also that although our analytical technique appears to be novel, the underlying mathematics of Bernoulli estimation is very conventional and can be found in any introduction to Bayesian inference.

In the analysis that follows we assume that each subject has separate Beta distributions Beta(α_v, β_v) and Beta(α_s, β_s) for the reliability of the visual and spoken cues respectively. For each subject for each maze, we estimated the four parameters $\alpha_v, \beta_v, \alpha_s$ and β_s via maximum likelihood (now including all trials, discordant and concordant). That is, the fitted values of these parameter represent the priors that best explain the subject's pattern of responses. Below we plot these estimates in (α, β) space (visual parameters in blue, spoken in red) which allows us to see the trust subjects placed in each source. Figs. 5 shows the modeling results for Exps. 1–3.

Note we fitted each subject and each maze individually, and then aggregated over subjects by estimating the distribution over observers in (α, β)-space, which is shown in the plots.

As can be seen by the positions of the peaks, in Exps. 1 and 3 the spoken cue shows substantial bias towards "trust" ($\alpha > \beta$) while the visual cue is closer to neutral ($\alpha = \beta$). The bias is mostly absent in the density for Exp. 2 (Fig. 5b) except for one cluster of estimates biased towards trusting the spoken cue, consistent with our finding of a small non-significant trend in this condition. Overall the results confirm the conclusions we drew above, but quantifies the nature and distribution of the spoken bias more precisely.

Each of the plots shows some evidence of multimodality in the density in (α, β) space. This implies the presence of potentially distinct decision-making strategies. The Exp. 1 results, for example (Fig. 5a), suggest that most subjects interpret the spoken cue in manner biased towards trust (a large cluster below the $\alpha = \beta$ diagonal), while subjects biases towards the visual cue are divided into two clusters, a large one approximately neutral (on the diagonal) and a smaller one strongly biased to trust (near $\alpha = 22, \beta = 3$). Similarly while the results of Exp. 3 broadly corroborated the bias towards spoken cues found in Exp. 1, the density plots suggest that most subjects were neutral to spoken cues (large cluster on the diagonal) while the effect was created by a smaller cluster biased towards trusting the spoken cue (at about $\alpha = 30, \beta = 10$)—while virtually none of the subjects were biased to the visual cue. It would be interesting to correlate such individual differences with other measurable personality or demographic factors, but our existing dataset does not allow such an analysis.

Discussion

The main conclusion from our results is that when integrating visual and spoken cues in a navigation task, subjects place a higher a priori weight on spoken instructions. This cannot be explained by the objective statistical evidence available to them, which was equated between cue types. Instead it seems to constitute a subjective bias on the part of (at least some of the) subjects through which they interpret evidence. To be sure, this does not mean their bias is "irrational;" indeed our modeling shows that it can be well-accounted for by a rational (i.e. Bayesian) inference process with Beta prior biased towards believing that spoken sources usually emit correct instructions ($\alpha_s - \beta_s > \alpha_v - \beta_v$). In the real world, of course, such a prior might be entirely reasonable.

Nominally, this main conclusion contradicts a number of previous multimodal studies suggesting that visual cues are generally weighted more heavily (e.g. Ghahramani, Wolptrt, & Jordan, 1997; Battaglia, Jacobs, & Aslin, 2003; Ernst & Bülthoff, 2004; Alais & Burr, 2004). However our auditory cues, unlike those in the relevant literature, comprised linguistic instructions from an apparently human source, and presumably the disparity derives in some way from this difference. Indeed this conclusion is bolstered by the finding that the bias in favor of spoken cues was diminished or even eliminated when the instructions were linguistically incomplete or ill-formed, as in Exp. 2.

Spoken instructions differ in a number of ways from nonverbal auditory cues. First, spoken communication is situated in a social context given special status in decision-making (Mesoudi, Whiten, & Dunbar, 2006; Rendell et al., 2011). Listeners assume that speakers provide information that is truthful and relevant to the situation (Grice, 1989). Even 17-month-old human infants expect other humans (at least in-group members) to provide helpful assistance (Jin & Baillargeon, 2017). Of course, the visual cue might also have originated from a human source, but the spoken cue apparently makes this connection more subjectively salient. Moreover, as our subjects were neurotypical adults with presumably intact theory of mind, they may have expected the speaker to have knowledge of the maze that they themselves lacked. In this sense, the priority they placed on spoken instructions may reflect known interactions between language and theory of mind (de Villiers, 2007). This hypothesis could be tested more directly by a manipulation in which the speaker's ability to personally see the hidden parts of the maze was somehow conveyed to the subject. In the absence of such a manipulation we can only infer that our subjects generally assumed that the speaker had up-to-date information, and meant to accurately convey it to the listener. It is also unclear whether the weight subjects placed on spoken cues reflected their *verbal* nature or their *auditory* nature. These could be deconfounded by using linguistic instructions that are not auditory (i.e. written signage) and auditory instructions that are not linguistic (e.g. a beep indicating direction).

In summary our main finding of a priority given to spoken instructions may reflect previously-known biases—but note that these biases have not generally been quantified in as precise a manner as our Bayesian framework allows. More broadly, these results point to the need to enlarge the study of cue integration in decision making to include social and linguistic contexts which are ubiquitous in human interaction.

Acknowledgments

We are grateful to Eileen Kowler for many discussions, and to Shahan Akhter for assistance in developing the Oculus Rift / Unity interface.

References

Abrams, R. A., & Christ, S. E. (2003). Motion Onset Captures Attention. *Psychological Science*, *14*(5), 427-432. (00415) doi: 10.1111/1467-9280.01458

Alais, D., & Burr, D. (2004). The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology*, *14*(3), 257-262. (01412) doi: 10.1016/j.cub.2004.01.029

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419–439.

Andersen, R. A., Snyder, L. H., Bradley, D. C., & Xing, J. (1997). Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience*, *20*(1), 303-330.

Backus, B. T. (2009). The mixture of bernoulli experts: A theory to quantify reliance on cues in dichotomous perceptual decisions. , *9*(1), 6–6. (00026) doi: 10.1167/9.1.6

Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *JOSA A*, *20*(7), 1391–1397. (00402)

Boersma, P., & Weenink, D. (2018). *Praat: doing phonetics by computer.* (Version 6.0.16, retrieved March 2018 from http://www.praat.org/)

Burr, D., Banks, M. S., & Morrone, M. C. (2009). Auditory dominance over vision in the perception of interval duration. *Experimental Brain Research*, *198*(1), 49. (00119) doi: 10.1007/s00221-009-1933-z

Cheng, K., Shettleworth, S. J., Huttenlocher, J., & Rieser, J. J. (2007). Bayesian integration of spatial information. *Psychological Bulletin*, *133*(4), 625-637. (00255) doi: 10.1037/0033-2909.133.4.625

Colavita, F. B. (1974, March). Human sensory dominance. *Perception & Psychophysics*, *16*(2), 409-412. (00488) doi: 10.3758/BF03203962

de Villiers, J. G. (2007). The interface of language and theory of mind. *Lingua*, *117*(11), 1858–1878.

Deneve, S., & Pouget, A. (2004). Bayesian multisensory integration and cross-modal spatial links. *Journal of Physiology-Paris*, *98*(1-3), 249-258. (00210) doi: 10.1016/j.jphysparis.2004.03.011

Durrant-Whyte, H., & Bailey, T. (2006). Simultaneous localization and mapping: part I. *IEEE Robotics Automation Magazine*, *13*(2), 99-110. doi: 10.1109/MRA.2006.1638022

Egeth, H. E., & Sager, L. C. (1977, January). On the locus of visual dominance. *Perception & Psychophysics*, *22*(1), 77-86. (00087) doi: 10.3758/BF03206083

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429-433. (03331) doi: 10.1038/415429a

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. , *8*(4), 162–169. (01576) doi: 10.1016/j.tics.2004.02.002

Farroni, T., Johnson, M. H., Brockbank, M., & Simion, F. (2000). Infants' use of gaze direction to cue attention: The importance of perceived motion. *Visual Cognition*, *7*(6), 705-718. (00218) doi: 10.1080/13506280050144399

Ghahramani, Z., Wolptrt, D. M., & Jordan, M. I. (1997). Computational models of sensorimotor integration. In *Advances in psychology* (Vol. 119, pp. 117–147). Elsevier. doi: 10.1016/S0166-4115(97)80006-4

Golledge, R. (2003). Human wayfinding and cognitive maps. In *The Colonization of Unfamiliar Landscapes* (p. 49-54). Routledge. (00908)

Grahn, J. A., Henry, M. J., & McAuley, J. D. (2011). FMRI investigation of cross-modal interactions in beat perception: Audition primes vision, but not vice versa. *NeuroImage*, *54*(2), 1231-1243. doi: 10.1016/j.neuroimage.2010.09.033

Grice, P. (1989). *Studies in the way of words.* Cambridge, MA: Harvard University Press.

Griffiths, T., Kemp, C., & Tenenbaum, J. (2008). Bayesian models of cognition. In *The Cambridge Handbook of Computational Psychology.* (00417)

Holland, S., Morse, D. R., & Gedenryd, H. (2002). AudioGPS: Spatial audio navigation with a minimal attention interface. *Personal and Ubiquitous computing*, *6*(4), 253-259. (00330)

Jin, K.-S., & Baillargeon, R. (2017). Infants possess an abstract expectation of ingroup support. *Proceedings of the National Academy of Sciences*, *114*(31), 8199–8204.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Research*, *35*(3), 389–412.

Lindley, D. V., & Phillips, L. D. (1976). Inference for a Bernoulli process (a Bayesian view). *The American Statistician*, *30*(3), 112–119.

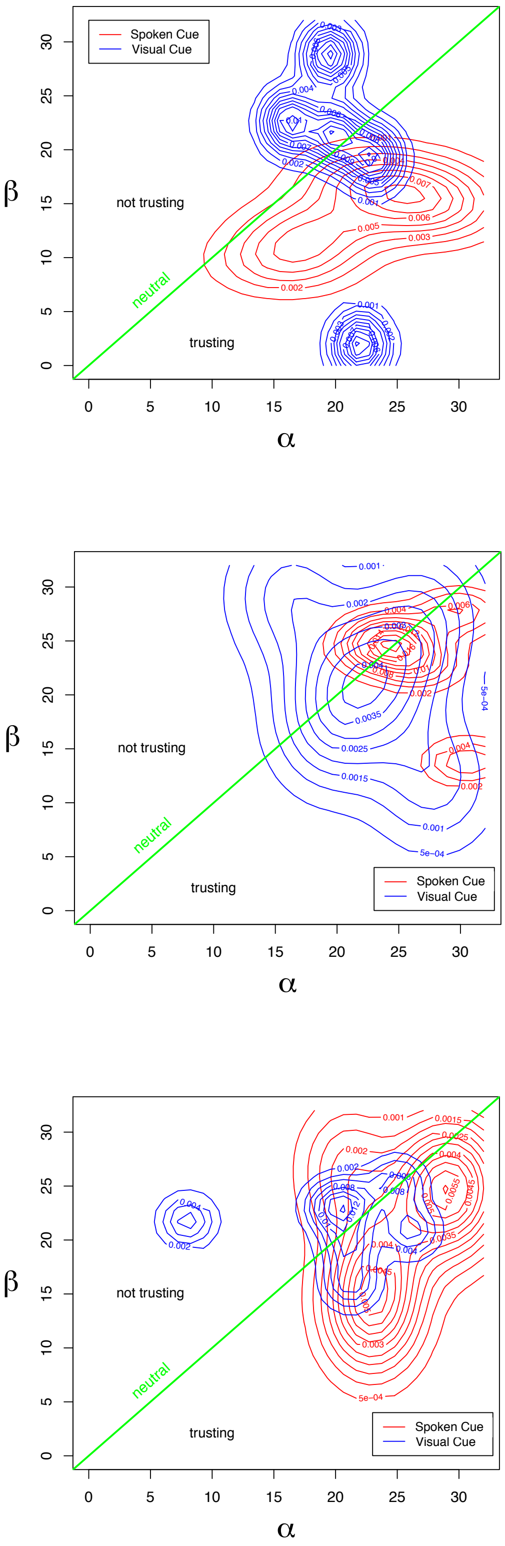
Loomis, J. M., Klatzky, R. L., Golledge, R. G., & Philbeck, J. W. (1999). Human navigation by path integration. *Wayfinding behavior: Cognitive mapping and other spatial processes*, 125–151.

Mesoudi, A., Whiten, A., & Dunbar, R. (2006). A bias for social information in human cultural transmission. *British Journal of Psychology*, *97*(3), 405-423. (00242) doi: 10.1348/000712605X85871

Rendell, L., Fogarty, L., Hoppitt, W. J., Morgan, T. J., Webster, M. M., & Laland, K. N. (2011). Cognitive culture: Theoretical and empirical insights into social learning strategies. *Trends in Cognitive Sciences*, *15*(2), 68-76. (00311) doi: 10.1016/j.tics.2010.12.002

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default bayes factors for ANOVA designs. , *56*(5), 356–374. (00978) doi: 10.1016/j.jmp.2012.08.001

Yuille, A. L., & Bülthoff, H. H. (1996). Bayesian decision theory and psychophysics. In D. C. Knill & W. Richards (Eds.), *Perception as Bayesian inference* (pp. 123–162). Cambridge: Cambridge University Press.



(a)

(b)

(c)

Figure 5: Contour plots of density estimates of α and β parameters in (a) Exp. 1 (b) Exp. 2 and (c) (a) Exp. 3. Visual cue parameters are in blue, spoken in red. Probability density below the line (α > β) connotes a “trust” in the cue source (a prior favoring believing it is usually reliable); above the line (α < β) connotes “distrust”; and the diagonal (α = β) indicates neutrality.