Current Biology, Vol. 14, 000–000, March 23, 2004, ©2004 Elsevier Science Ltd. All rights reserved. DOI 10.1016/j.cub.2004.xx.xxx

Neuroscience: What You See and Hear is What You Get

Dispatch

Martin S. Banks

The brain receives signals from a variety of sources; for example, visual and auditory signals can both indicate the direction of a stimulus, but with differing precision. A recent study has shed light on the way that the brain combines these signals to achieve the best estimate possible.

You enter a crowded room and someone calls your name. You turn to see who it is. You now see several people in the general direction the voice came from. Many are talking. Which one called your name? You hear it again and now the sound seems to come from straight ahead or nearly so. There are still a handful of candidates in your field of view, so you look from one to the other. Finally, you see one whose lips move as you hear your name once more. Sound and sight have come together and you identify the speaker as your college roommate. How does this work? That is, how does the brain find the appropriate auditory–visual correspondence to determine that a sound and sight have come from the same source? In a study published recently in *Current Biology*, Alais and Burr [1] have demonstrated an important and seemingly pervasive rule for the combination of visual and auditory cues to spatial location.

Before discussing their experiment, it is useful to describe what we know about visual and auditory localization. The direction of stimulus can be represented by two coordinates: the azimuth (horizontal direction in angular units) and the elevation (vertical). The eye is well suited for determining direction, because the direction in which a light ray enters the eye is directly indicated by the position it stimulates on the retina. As a consequence, the visual system can distinguish very small directional changes. The 'just-noticeable difference' (JND) in the position of one small stimulus relative to another is roughly 5-10 seconds of arc [2].

The auditory system is not nearly so well suited for determining direction, because the direction a sound enters the ear must be calculated from a variety of cues. One set of cues arises from the spatial separation of the two ears. Because of the separation, most sounds travel different distances to the left and right ears. A sound on the left, for example, reaches the left ear slightly before reaching the right ear, thereby creating an interaural time difference. Similarly, interaural intensity differences arise because the head acts as a sound shadow. A sound on the left is attenuated more on its way to the right ear than on its way to the left ear. The azimuth of a sound source can be determined from these interaural time and intensity differences, but the JND is relatively large, typically 1–2 degrees [3].

It is unclear how interaural differences alone can signal changes in elevation, yet people can reliably distinguish elevation changes of 5–10 degrees [4]. Batteau [5] showed how the filtering effect of the outer ear (pinna) can provide the missing information. Specifically, the spectrum of an approaching sound is filtered by interactions with the groves and ridges of the pinna, and the listener is able to use the filtered spectrum to judge elevation.

The visual system is thus far better suited than the auditory system for estimating direction. In Alais and Burr's [1] study, the visual JND was 5–10 times lower (more precise) than the auditory JND. It is widely believed that such differences in the precision of localization lead to visual capture, in which the apparent direction of an auditory stimulus is determined largely by the direction of a corresponding visual stimulus. Ventriloquism is an entertaining example of this phenomenon [6]. Alais and Burr [1] asked whether visual capture of a sound's apparent direction derives from a rigid rule, in which the visual estimate always determines the overall percept, or from a more general procedure of weighting sensory evidence in a statistically optimal fashion.

What would the best way be to combine noisy — variable — auditory and visual estimates of direction? The answer depends on the goal. If one wants an estimate that is unbiased and has minimum variability, the best combination rule in most cases is a weighted average. Suppose the brain has unbiased estimates D_A and D_V based on auditory and visual signals, respectively. Those estimates will have moment-to-moment variability because of fluctuations in the physical stimuli and noisiness in the brain's measurements of them. The variability of D_A and D_V can be represented by variances σ_A^2 and σ_V^2. For convenience, define the reliabilities of D_A and D_V as reciprocal variances, $r_A = 1/\sigma_A^2$ and $r_V = 1/\sigma_V^2$. Under reasonable assumptions, the rule yielding the lowest-variance, unbiased estimate is a weighted average [7]:

$$D = w_A D_A + w_V D_V \qquad (1)$$

$$w_A = r_A/(r_A + r_V) \text{ and } w_V = r_V/(r_A + r_V)$$

These equations are derivable from Bayes' Law, the statistical rule that prescribes how to take evidence and potential costs into account when making a decision [8]. The reliability of the resulting estimate is:

$$r = r_A + r_V \qquad (2)$$

Vision Science Program, School of Optometry, Department of Psychology, and Wills Neuroscience Institute, University of California, Berkeley California 94720 USA.
E-mail: marty@john.berkeley.edu

Thus, the reliability resulting from this weighted average will always be greater than the reliability of either of the sensory estimates alone. Said another way, this combination rule yields an estimate of lowest possible variance.

Alais and Burr [1] investigated whether people follow this rule when estimating the direction of an auditory-visual stimulus. They first conducted single-cue experiments to measure reliabilities for purely auditory and purely visual stimuli. Those experiments yielded estimates of the variances σ_A^2 and σ_V^2 which were then used to specify the weights in equation (1). They then conducted a two-cue experiment with two kinds of stimulus: a non-conflict stimulus with visual and auditory stimuli in the same direction; and a conflict stimulus with the two in slightly different directions.

The middle panel of Figure 1 shows representative distributions for the conflict stimulus. The auditory part of the stimulus is presented at –6 degrees relative to straight ahead (green vertical arrow) and the visual part at –2 degrees (yellow arrow). The green and yellow curves D_A and D_V represent probability distributions of the estimated directions of the auditory and visual parts of the stimulus. If the brain uses the optimal combination rule — equation (1) — then the distribution of resulting estimates would be the white curve D, which peaks at a value closer to D_V than D_A, and has lower variance than either.

If the variance of D_V were much lower than the variance of D_A, the white and yellow curves would nearly superimpose, so it would be difficult to determine whether the brain was using just the visual signal or was averaging the visual signal with high weight and the auditory signal with low weight. Alais and Burr [1] circumvented this problem by blurring the visual stimulus in some cases to make its direction uncertain. The optimal rule predicts that, in these cases, the auditory stimulus will mostly determine perceived direction ('auditory capture').

Alais and Burr [1] found clear evidence for visual dominance when the visual stimulus was sharply focused, and for auditory dominance when it was blurred. In both cases, the observed percepts were very close to those predicted by the optimal combination rule. Furthermore, when auditory and visual stimuli were both present, subjects made finer direction discriminations than from either sense alone, again as predicted by the optimal rule. These observations are consistent with the hypothesis that the brain uses an optimal combination rule, based on the relative reliabilities of sensory inputs, to determine perceived direction. Their results add to a growing consensus that a statistically optimal (or nearly optimal) combination rule is used for combining signals from different senses [9–13] and for combining cues within a sense [11,14,15].

The observation of optimal or nearly optimal cue combination points to two vexing questions. First, how does the brain know the variances of its sensory estimates in order to make correct weight assignments? Recent work indicates how this could be achieved using population codes [9,16,17]. And second, how does the brain know when sensory estimates are coming from the same source and not

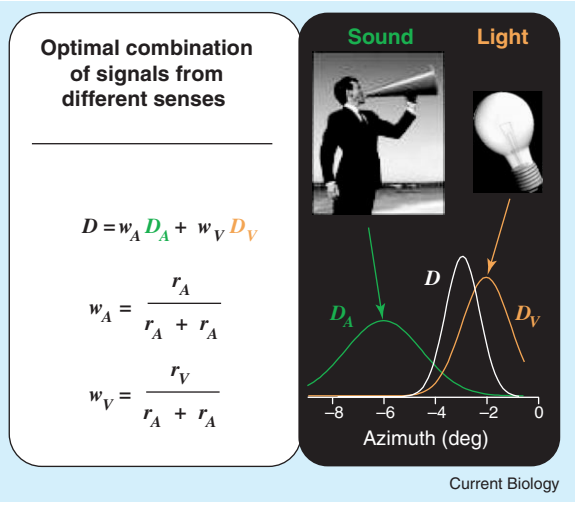

Figure 1. Optimal combination of signals from different senses.

The green and yellow curves represent the probability distributions associated with an auditory stimulus presented at an azimuth of –6 degrees (green vertical arrow) and a visual stimulus presented at –2 degrees (yellow arrow). The white curve represents the distribution that would result by using the optimal combination rule (equation (1) in the text). The peak is closer to the visual than to the auditory distribution and the variance of the combined distribution is less than the variances of the other two distributions. Alais and Burr [1] presented two types of auditory–visual stimulus: a non-conflict stimulus, in which the auditory and visual directions were identical; and a conflict stimulus, in which the two differed slightly. They varied the direction of the non-conflict stimulus to find the point of subjective equality (PSE): the value that on average had the same perceived direction as the conflict stimulus. In the figure, the perceived direction of the conflict stimulus should be roughly –3 degree, so if the brain uses the weighted-average rule, the PSE would be at –3 degrees.

different sources, so that combining makes sense? Neurophysiological work in the brainstem and cortex has revealed circuits that might be involved in handling this problem of inter-cue correspondence [18]. Obtaining answers to these questions is an important future challenge for neuroscientists and perceptual psychologists.

There are also practical benefits to the study of cue combination. More and more applications are being found for 'virtual reality', including remote devices (such as tele-surgery), scientific visualization, education and training (for example, surgical training), computer-aided design and virtual prototyping, and entertainment. In addition to the standard three-dimensional visual simulations, virtual reality systems are now adding haptic and tactile displays and three-dimensional audio displays to improve realism and usefulness. Knowing the combination rules employed by typical human operators will allow virtual reality engineers to make more informed choices about the precision requirements for the various senses. As multi-sensory virtual reality becomes more effective and commonplace, you may someday see and hear a convincing simulation of your college roommate.

References

1. Alais, D., and Burr, D. (2004). The ventriloquist effect results from near

optimal crossmodal integration. Curr. Biol., February 3 issue.

2. Westheimer, G., and McKee, S.P. (1977). Spatial configurations for visual hyperacuity. Vis. Res. *17*, 941-947.
3. Mills, A. (1958). On the minimum audible angle. J. Acoust. Soc. Am. *30*, 237-246.
4. Perrott, D., and Saberi, K. (1990). Minimum audible angle thresholds for sources varying in both elevation and azimuth. J. Acoust. Soc. Am. *87*, 1728-1731.
5. Batteau, D.W. (1967). The role of the pinna in human localization. Proc. Roy. Soc. B. *168*, 158-180.
6. Connor, S. (2000). Dumbstruck: A Cultural History of Ventriloquism. (Oxford: Oxford University Press).
7. Cochran, W.G. (1937). Problems arising in the analysis of a series of similar experiments. J. Royal Stat. Soc. *4 (suppl)*, 102-118.
8. Clarke, J.J., and Yuille, A.L. (1990). Data Fusion for Sensory Information Processing. (Boston: Kluwer Academic).
9. Ernst, M.O., and Banks, M.S. (2002). Human integrate visual and haptic information in a statistically optimal fashion. Nature *415*, 429-433.
10. van Beers, R.J., Wolpert, D.M., and Haggard, P. (2002). When feeling is more important than seeing in sensorimotor adaptation. Curr. Biol. *12*, 834-837.
11. Hillis, J.M., Ernst, M.O., Banks, M.S., and Landy, M.S. (2002). Combining sensory information: Mandatory fusion within, but not between, senses. Science *298*, 1627-1630.
12. Gepshtein, S., and Banks, M.S. (2003). Viewing geometry determines how vision and haptics combine in size perception. Curr. Biol. *13*, 483-488.
13. Körding, K.P., and Wolpert, D.M. (2004). Bayesian integration in sensorimotor learning Nature 427, 244-247.
14. Landy, M.S., and Kojima, H. (2001). Ideal cue combination for localizing texture defined edges. J. Opt. Soc. Am. A *18*, 2307-2320.
15. Knill, D.C., and Saunders, J.A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? Vis. Res. *43*, 2539-2558.
16. Pouget, A., Dayan, P., and Zemel, R.S. (2003). Computation and inference with population codes. Ann. Rev. Neurosci. *26*, 381-410.
17. Deneve, S., Latham, P.E., and Pouget, A. (1999). Reading population codes: a neural implementation of ideal observers. Nat. Neurosci. *2*, 740-745.
18. Stein, B.E., and Meredith, M.A. (1993). The Merging of the Senses. (Cambridge: MIT Press).