Current Biology, Vol. 14, 257–262, February 3, 2004, ©2004 Elsevier Science Ltd. All rights reserved. DOI 10.1016/j.cub.2004.01.029

The Ventriloquist Effect Results from Near-Optimal Bimodal Integration

David Alais[1,2] and David Burr[1,3,*]
[1]Istituto di Neuroscienze del CNR
56127 Pisa
Italy
[2]Auditory Neuroscience Laboratory
Department of Physiology
University of Sydney
New South Wales 2006
Australia
[3]Department of Psychology
University of Florence
50125 Florence
Italy

Summary

Ventriloquism is the ancient art of making one's voice appear to come from elsewhere, an art exploited by the Greek and Roman oracles, and possibly earlier [1]. We regularly experience the effect when watching television and movies, where the voices seem to emanate from the actors' lips rather than from the actual sound source. Originally, ventriloquism was explained by performers projecting sound to their puppets by special techniques [1], but more recently it is assumed that ventriloquism results from vision "capturing" sound [2–5]. In this study we investigate spatial localization of audio-visual stimuli. When visual localization is good, vision does indeed dominate and capture sound. However, for severely blurred visual stimuli (that are poorly localized), the reverse holds: sound captures vision. For less blurred stimuli, neither sense dominates and perception follows the mean position. Precision of bimodal localization is usually better than either the visual or the auditory unimodal presentation. All the results are well explained not by one sense capturing the other, but by a simple model of optimal combination of visual and auditory information.

Results and Discussion

Observers were required to localize in space brief light "blobs" or sound "clicks," presented either separately (unimodally) or together (bimodally). In a given trial, two sets of stimuli were presented successively (separated by a 500 ms pause) and observers were asked to indicate which appeared to be more to the left, guessing if unsure. The visual stimuli were low-contrast (10%) Gaussian blobs of various widths, back-projected for 15 ms onto a translucent perspex screen (80 × 110 cm) subtending 90° × 108° when viewed binocularly from 40 cm. The auditory stimuli were brief (1.5 ms) clicks presented though two visible high-quality speakers at the edge of the screen, with the apparent position of the sound controlled by interaural time differences.

*Correspondence: dave@in.cnr.it

Results for the various unimodal location discriminations for naive observer L.M. are shown in Figure 1A. The curves plot the proportion of trials in which the second stimulus was seen to the left of the first, as a function of actual physical displacement. Following standard practice, the data were fitted with cumulative Gaussian functions free to vary in position and width: the position of the median (50% leftward) is termed the point of subjective equality (PSE), and the width σ represents the estimate of localization accuracy (presumed to depend on internal noise). For all unimodal conditions, the PSE was near 0°, but σ varied considerably. For visual stimuli, σ was smallest (approximately 1°) for the smallest Gaussian blobs and largest (approximately 20°) for the largest (in line with the results of [6]). Localization accuracy for the auditory click was around 6°, falling midway between the various visual stimuli. Note that this poor localization does not necessarily reflect performance under natural or "free-field" conditions where all auditory cues (including intensity differences and monaural cues) are available. Free-field localization of spectrally rich stimuli such as click trains produces discrimination thresholds on the order of 1° [7, 8].

With unimodal thresholds established, we then measured localization for a blob and click presented simultaneously. Observers were asked to envisage each presentation as a single event, like a ball hitting the screen, and report which of the two presentations was more leftward. For one presentation (randomly first or second), the visual and auditory stimuli were in conflict, with the visual stimulus displaced Δ degrees rightward and the auditory stimulus displaced Δ degrees leftward ($S_V - S_A = 2\Delta$, where S_V and S_A are the spatial positions of the visual and auditory stimuli). The average position of this stimulus was always zero, as in previous studies [9]. On the other (non-conflict) presentation, the two modalities were covaried to the left or right of center by the amount shown in the abscissa (with positive meaning rightward).

The different colored symbols in Figures 1B–1D show results for $\Delta = \pm 5°$, $\pm 2.5°$, or 0°. The pattern of results depended critically on blob size, which increases through (B)–(D). For small blob widths (4°; Figure 1B), the curves were displaced systematically in the direction of Δ (to the side where the visual stimulus was displaced), suggesting that vision dominated the perceived position of the incongruent stimuli (the ventriloquist effect). However, for large blobs (64°; Figure 1D), the reverse held, with the curves displaced in the opposite direction of Δ, indicating that the click dominated perceived position, an "inverse ventriloquist effect." For the mid-sized blobs (32°; Figure 1C), the curves tended to cluster together, suggesting that the perceived position depended on an average of the two modalities.

Figure 2 summarizes the results of Figure 1, together with those of two other observers, S.D. (naive) and author D.A. The upper panels show the PSE, the median displacement of the conflict stimulus (calculated from

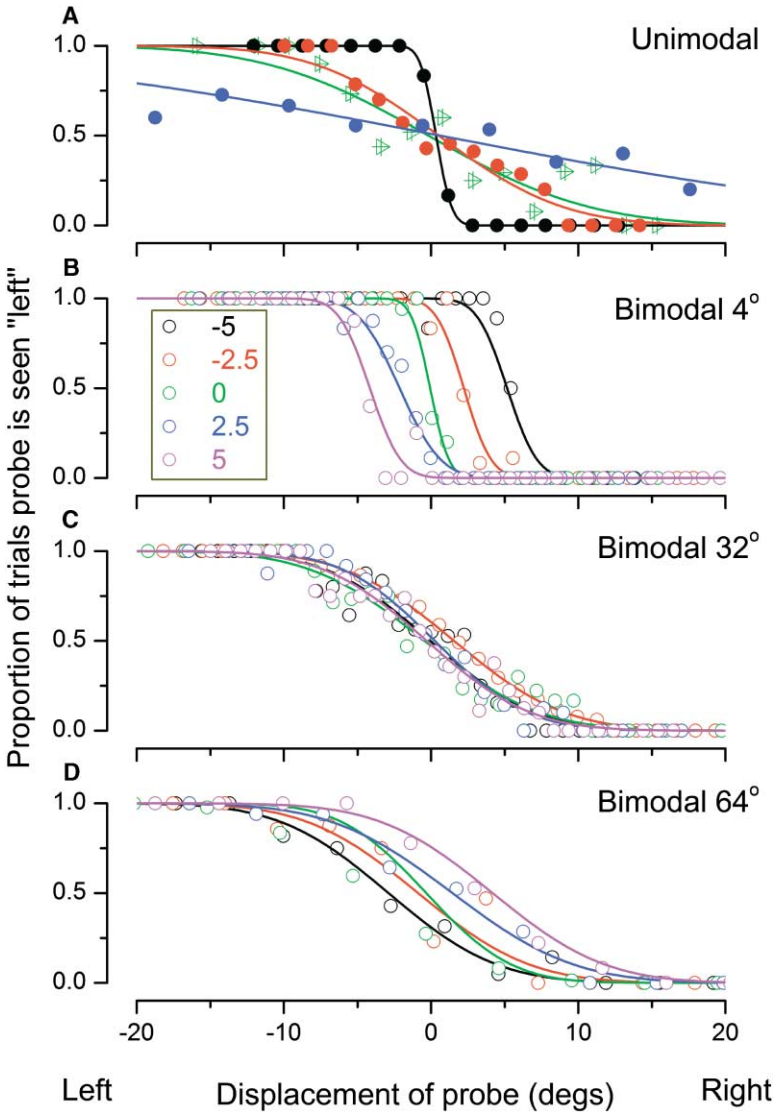

Figure 1. Unimodal and Bimodal Localization of Visual and Auditory Stimuli

Results are for one observer L.M., naive of the goals and actual conditions of the experiment. Four other subjects (two summarized in Figure 2 and two not reported) behaved similarly.

(A) Psychometric functions for localizing either an auditory click (green speaker-shaped symbols) or visual blobs of various Gaussian space constants ($2\sigma = 4°$, black; $2\sigma = 32°$, red; or $2\sigma = 64°$, blue). The auditory stimuli were brief (1.5 ms) clicks, with their apparent position on the median plane controlled by interaural time differences (ITDs). Timing resolution was 15.3 μs, corresponding to about 1.2° in lateral angle (calibrated separately for each observer). All trials comprised two stimulus presentations, one presented near-centrally (with a small jitter from trial to trial) and the other displaced leftward or rightward by an amount given by the abscissa. The ordinate shows the proportion of times the observer judged the probe presentation (randomly first or second) "leftward." Two hundred fifty trials were run for each condition, over ten separate sessions. The data show that the points of subjective alignment are similar for all stimuli (≈0°), while the widths of the visual functions (assumed to reflect internal neural noise) increase with the width of the visual stimulus (see also Figure 2B). The width of the auditory function lies midway between the smallest and largest visual stimuli.

(B–D) Psychometric functions for localizing bimodal presentations of the click and blob together (click centered within the blob), for blob widths 4° (B), 32° (C), or 64° (D). One presentation (randomly first or second) was the conflict stimulus, with the visual stimulus horizontally displaced Δ° rightward and the auditory stimulus displaced Δ° leftward. In the other non-conflict presentation, both stimuli were displaced in the same direction by the amount shown by the abscissa (positive indicates rightward). The values of Δ were −5° (black symbols), −2.5° (red), 0° (green), +2.5° (blue), or +5° (mauve), randomly intermingled within each session. Two hundred fifty trials were run for each condition, over ten separate sessions. Figure 2 summarizes these data, together with those of another two subjects.

Gaussian fits like those in Figure 1), as a function of the audio-visual conflict (Δ). The blue and black dashed lines (respectively) show the predicted results if vision or audition were to dominate completely. The continuous lines are model predictions (described below), in close agreement with the data. As noted previously, for small blobs (black symbols), the PSE varies directly with Δ; for large blobs (blue symbols), it varies inversely; and for mid-sized blobs (red and green symbols), it was almost independent of Δ. The lower panels show the localization-accuracy thresholds for the various conditions, given by the standard deviations of the Gaussian fits. Visual accuracy varied directly with the width of the Gaussian blobs, while localization accuracy for the auditory click was similar to those of the mid-sized blob. The combined thresholds (average of all the conflict conditions), indicated by the red circles, are usually lower than either visual or auditory threshold alone.

Several authors [9–13] have recently suggested that multimodal information may be combined in an optimal way by summing the independent stimulus estimates from each modality according to an appropriate weighting scheme. The weights correspond to the inverse of the noise associated with each estimate, given by the variance σ^2 of the underlying noise distribution (assumed to be approximated by the squared width of the psychometric function). This model is "optimal" in that it combines the unimodal information to produce a multimodal stimulus estimate with the lowest possible variance (that is, with the greatest reliability). Optimum combination of the independent auditory and visual estimates $\hat{S}_A$ and $\hat{S}_V$ is given by (see [10]):

$$\hat{S}_{VA} = w_V\hat{S}_V + w_A\hat{S}_A \quad (1)$$

where w_V and w_A are the relative weights for each modality, inversely proportional to their localization variances:

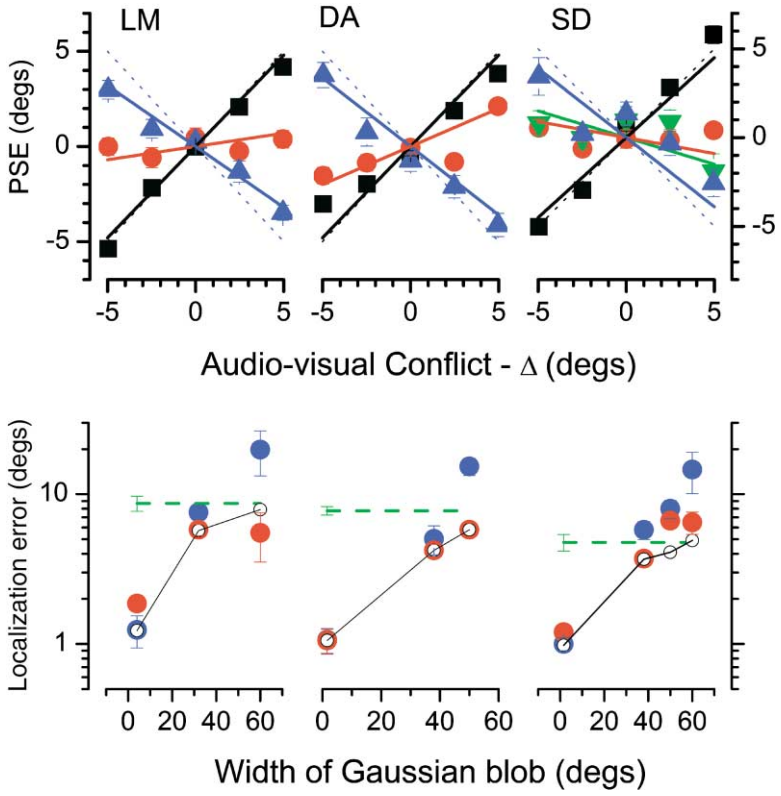

Figure 2. Effect of Conflict on Localization of Visual and Auditory Stimuli

(A) PSE for localizing the bimodal stimulus as a function of audio-visual conflict (Δ). The PSE was defined as the median of the cumulative Gaussian distributions for each observer (see examples of Figure 1). The results are shown for various sizes of visual blob, indicated by the colors (L.M.: 4°, black squares; 32°, red circles; and 64°, blue triangles; D.A.: 1.6°, 38°, and 50°; S.D.: 1.6°, 38°, and 50°, green triangles; and 64°). For the small blob sizes, the perceived bimodal position varied directly with the visual stimulus, while for the larger blobs they varied with audition. For the mid-sized blobs, there is little variation. The dotted black and blue lines show the predicted result if vision and audition were to dominate totally. The continuous lines are not fits to the data, but predictions from optimal statistical combination of cues (Equation 1), obtained solely from the parameters of the unimodal measurements (shown in [B]).

(B) Localization error (given by the root-variance of the psychometric functions) as a function of blob size. The blue symbols show the unimodal visual thresholds and the green dashed line the auditory unimodal thresholds. The bimodally measured thresholds are shown by the filled red circles, and the predicted thresholds (from optimal combination: Equation 3) by the black open circles joined by the lines. Note that for all subjects, the second point along the abscissa is where visual and auditory localization errors are most similar. In these cases, the model predicts that bimodal variance should be lower than those of either vision or audition (see also Figure 3).

$$w_A = \frac{1/\sigma_A^2}{1/\sigma_A^2 + 1/\sigma_V^2} = \frac{\sigma_V^2}{\sigma_A^2 + \sigma_V^2} \quad (2)$$

and likewise for W_V. Estimates of σ_V^2 and σ_A^2 can be obtained by the Gaussian fit of the unimodal data of Figure 1A (also plotted in Figure 2B). From these we calculated visual and auditory weights for the various blob widths and predicted how PSE should vary with Δ, assuming that the best visual and auditory estimates ($\hat{S}_V$ and $\hat{S}_A$) were given by the actual position of the visual and auditory sources, $-\Delta$ and Δ, respectively (given that the unimodal location estimates were all close to zero). These predictions are shown by the continuous lines in Figure 2A, and clearly fall very close to the actual data. Note that the predictions have no free parameters and are based solely on the unimodal data, not the data of Figure 2A.

An important further prediction of optimal combination of visual and auditory information is that the variance of the combined estimate will always be less than either individual estimate, provided that the underlying noise distributions are independent:

$$\sigma_{VA}^2 = \frac{\sigma_V^2\sigma_A^2}{\sigma_A^2 + \sigma_V^2} \leq \min(\sigma_V^2, \sigma_A^2). \quad (3)$$

The reduction in combined variance will be greatest when $\sigma_V^2 = \sigma_A^2$ and least when they are most different (where the lower variance will dominate). The model does not predict a large improvement in localization accuracy, $\sqrt{2}$ at best, but the prediction does provide a further test of the optimal combination model. If a reduction in variance is observed, it shows that the combined estimate of location is not merely an averaging of the two unimodal estimates, as might occur by sampling randomly from each. The black lines in Figure 2B show the predicted thresholds from the optimal combination. For the smallest and largest blob sizes, the difference between the optimal combination prediction and the best unimodal estimate (visual or auditory) is very small and virtually impossible to assess statistically. However, for blob sizes around 30°, where σ_V and σ_A are similar, there is a small improvement predicted in localization accuracy which approaches $\sqrt{2}$. For the subjects in Figure 2B, the predicted improvement ranges from 1.18 to 1.35 (since the unimodal variances are not exactly identical, they do not quite reach $\sqrt{2}$). The results tend to follow the predictions, but not perfectly. While for D.A. (the most experienced psychophysical observer of bimodal stimuli) the predictions are very good indeed, for S.D. the combined performance is worse than predicted for the two largest blurs, and for L.M. for the smallest blur. While performance is in some cases not as predicted, there is only one statistically significant deviation from the predicted value (L.M., 64° blob, $p = 0.04$).

Figure 3 shows auditory, visual, bimodal, and predicted bimodal thresholds for one blob size (the one yielding the most similar auditory and visual thresholds and hence the largest predicted improvement) for each subject, and for the group mean. In all six subjects,

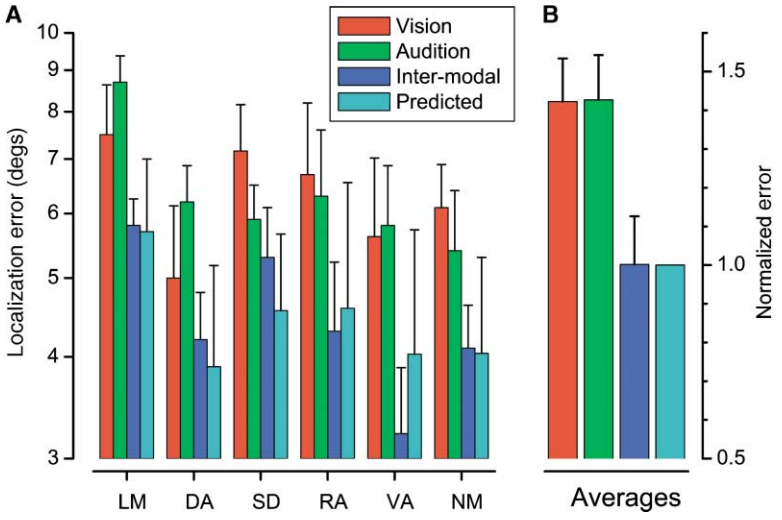

Figure 3. Comparison of Actual and Predicted Thresholds for Visuo-Auditory Localization

(A) Visual, auditory, bimodal, and predicted localization errors for six subjects, for the visual condition which yielded thresholds most similar to the auditory thresholds. Visual blobs of 38° were used in all cases except for L.M., in which they were 32°. This graph includes data from the three subjects of Figure 2B. The bars show standard errors of the mean, calculated by 500 repetitions of a bootstrap procedure [27], with the error of the predicted result calculated from the two unimodal thresholds, with propagation of errors based on Equation 3. The probabilities of rejecting the null hypothesis (that the bimodal threshold was the same as the best unimodal threshold), calculated by the bootstrap t test (one-tailed, 5000 repetitions) for the six observers were: L.M., 0.24; D.A., 0.5; S.D., 0.11; R.A., 0.057; V.A., 0.03; and N.M., 0.07.

(B) Averages of all six subjects, after normalizing to unity for the predicted value (so the error in that condition is zero). The improvement of the bimodal condition over the average unimodal conditions was 1.425, compared with a predicted value of 1.414. Averaging the performance of the six observers and comparing best unimodal performance with bimodal performance in a one-tailed, matched samples t test produced a clearly significant difference, with $t_5 = 3.69$ and $p < 0.01$. When the bimodal thresholds were tested against the predicted thresholds by the bootstrap t test technique, none were significantly different (all p values greater than 0.2).

bimodal localization was better than either unimodal localization and always near the predicted threshold. The predicted differences are not large compared with the measurement errors associated with estimating curve slopes (obtained from a bootstrap technique [27] and indicated by the error bars). Individually, on a paired t test analysis between the bimodal and the best unimodal condition, only two out of six cases reached 5% significance (details in caption). However, for none of the subjects is the bimodal threshold greater than the best unimodal estimate, and none differs statistically from the predicted threshold ($p > 0.2$ for all). Averaging the data, however, produces a clear result: the bimodal thresholds are lower than the average of the visual and auditory thresholds by a factor of 1.425 (compared with a predicted improvement of 1.414). Comparison of the best unimodal performance with bimodal performance in a one-tailed, matched samples t test produces a clearly significant result, with $t_5 = 3.69$ and $p < 0.01$.

Overall, the results support a model in which visual and auditory information is combined by minimizing variance, leading to an improvement in discriminating bimodal spatial location. The lack of statistical significance on individual subjects may be due to several factors. First, estimates of the width of psychometric functions are intrinsically variable, so measurement error can obscure true differences. Second, there may be an additional noise source at the level of bimodal combination, not considered in the model. Third, there may be correlations between the noise sources of the visual and auditory modalities, as has been demonstrated recently in other modalities [14, 15]. Fourth, factors such as attention and learning to use normally less-reliable modalities (sound in this case) should also be studied, particularly for untrained, naive observers.

We conclude that the ventriloquist effect is a specific example of near-optimal combination of visual and auditory space cues, where each cue is weighted by an inverse estimate of noisiness, rather than one modality capturing the other. As visual localization is usually far superior to auditory location, vision normally dominates, apparently capturing the sound source and giving rise to the classic ventriloquist effect. However, if the visual estimate is corrupted sufficiently by blurring the visual target over a large region of space, the visual estimate can become worse than the auditory one, and optimal localization correctly predicts that sound will effectively capture sight. This is broadly consistent with other reports of integration of sensory information [9–12, 16]. However, it differs slightly from the results of Battaglia et al. [13], who investigated localization in depth with visual and auditory cues in conflict, degrading visual performance by the introduction of noise (following [9]). They found that vision tended to dominate more than predicted by Equation 1, and were forced to introduce a hybrid Bayesian model to explain their effects. Unfortunately, they do not report thresholds for the bimodal conditions to determine whether bimodal presentation improved localization under these conditions, the more stringent test for statistically optimum combination. It is unclear why the results of this study are different from ours, but we can point to several differences in experimental design: Battaglia et al. required subjects to localize in depth (rather than in the horizontal plane), corrupted their visual stimuli with noise (rather than blurring), and presented both conflict and comparison stimulus simultaneously (rather than successively), possibly overloading the auditory system. More crucially, the conflict situation was always in the same direction, vision negative and audition positive, that could easily lead to "recalibration" (adaptation), considerably affecting the results [17, 18]. Finally, we cannot exclude the effects of learning and instructions: our subjects, both authors and naive, were trained extensively on the auditory task and were asked to think of the display as a ball thudding onto the screen to ensure that they attended to

both visual and auditory aspects of the stimuli. We cannot be certain if any of these differences was crucial in obtaining the current results, but it does seem reasonable that subjects be encouraged to use and to trust their nonpreferred sense and that the perceptual system not be given the chance to recalibrate to a consistent audiovisual discrepancy.

In the present study, for auditory localization to be superior to vision, the visual targets needed to be blurred extensively, over about 60°, enough to blur most scenes beyond recognition. However, we should recall that the location of the audio stimulus was defined by only one cue (interaural timing difference) and was not time varying, so auditory localization was only about one-sixth as accurate as normal hearing [7, 8]. If the effect were to generalize to natural hearing conditions, then 10° blurring would probably be sufficient. This is still a gross visual distortion, explaining why the reverse ventriloquist effect is not often noticed for spatial events. There are cases, however, when it does become relevant, not so much for blurred as for ambiguous stimuli, such as when a teacher tries to make out which child in a large class was speaking.

There is one previously reported case where sound does capture vision; this is for temporal localization where a small continuous (and peripherally viewed) light source seems to pulse when viewed together with a pulsing sound source [19, 20]. Furthermore, the presence of the clicks do not only make the light appear to flash, but can actually improve performance on visual discrimination tasks [21, 22]. Although no model was offered to account for this phenomenon, it may well result from sound having far better temporal acuity than vision, resulting in the sound information being heavily weighted and appearing to capture the visual stimulus.

It has recently been shown that ventriloquism can occur without specifically attending to the visual stimulus [23]. However, under conditions of ventriloquism, where sounds seem to be displaced in the direction of a visual stimulus, auditory attention can be attracted by visual cues, away from the actual locus of the sound [24]. It would be interesting to ascertain whether attending selectively to vision or audition can influence the ventriloquist effect.

An important and difficult remaining question is how the nervous system "knows" the variances associated with individual estimates. Must it "learn" these weights from experience, or could a direct estimate of variance be obtained from neural activity of a population, for example by observing the spread of activation along a spatiotopic map? Previous studies have shown that observers can learn cue-integration strategies [25] and that the learning can be very rapid [26]. We can only guess at the neural mechanisms involved, but it is not implausible that the central nervous system encodes an estimate of measurement error along with every estimate of position, or other attributes [9].

Acknowledgments

We acknowledge support from the Human Frontiers Science Programme and the Italian Ministry of Universities and Research, and a European Commission Marie Curie Fellowship to D.A.

Received: November 14, 2003
Revised: December 23, 2003
Accepted: December 23, 2003
Published: February 3, 2004

References

1. Connor, S. (2000). Dumbstruck: A Cultural History of Ventriloquism. (Oxford: Oxford University Press).
2. Pick, H.L., Warren, D.H., and Hay, J.C. (1969). Sensory conflict in judgements of spatial direction. Percept. Psychophys. *6*, 203–205.
3. Warren, D.H., Welch, R.B., and McCarthy, T.J. (1981). The role of visual-auditory "compellingness" in the ventriloquism effect: implications for transitivity among the spatial senses. Percept. Psychophys. *30*, 557–564.
4. Mateeff, S., Hohnsbein, J., and Noack, T. (1985). Dynamic visual capture: apparent auditory motion induced by a moving visual target. Perception *14*, 721–727.
5. Caclin, A., Soto-Faraco, S., Kingstone, A., and Spence, C. (2002). Tactile "capture" of audition. Percept. Psychophys. *64*, 616–630.
6. Hess, R.F., and Hayes, A. (1994). The coding of spatial position by the human visual system: effects of spatial scale and retinal eccentricity. Vision Res. *34*, 625–643.
7. Mills, A. (1958). On the minimum audible angle. J. Acoust. Soc. Am. *30*, 237–246.
8. 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.
9. Ernst, M.O., and Banks, M.S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature *415*, 429–433.
10. Clarke, J.J., and Yuille, A.L. (1990). Data Fusion for Sensory Information Processing. (Boston: Kluwer Academic).
11. Ghahramani, Z., Wolpert, D.M., and Jordan, M.I. (1997). Computational models of sensorimotor integration. In Self-Organization, Computational Maps and Motor Control.P.G. Morasso, V. Sanguineti, eds. (Amsterdam: Elsevier Science Publishing), pp. 117–147.
12. Jacobs, R.A. (1999). Optimal integration of texture and motion cues to depth. Vision Res. *39*, 3621–3629.
13. Battaglia, P.W., Jacobs, R.A., and Aslin, R.N. (2003). Baysian integration of visual and auditory signals for spatial localization. J. Opt. Soc. Am. A. Opt. Image Sci. Vis. *20*, 1391–1397.
14. Gepshtein, S., and Banks, M.S. (2003). Viewing geometry determines how vision and haptics combine in size perception. Curr. Biol. *13*, 483–488.
15. Oruc, I., Maloney, L.T., and Landy, M.S. (2003). Weighted linear cue combination with possibly correlated error. Vision Res. *43*, 2451–2468.
16. Alais, D., and Burr, D. (2003). No direction-specific bimodal facilitation for audiovisual motion detection. Brain Res. Cogn. Brain Res., in press.
17. Canon, L.K. (1970). Intermodality inconsistency of input and directed attention as determinants of the nature of adaptation. J. Exp. Psychol. *84*, 141–147.
18. Radeau, M., and Bertelson, P. (1974). The after-effects of ventriloquism. Q. J. Exp. Psychol. *23*, 63–71.
19. Shams, L., Kamitani, Y., and Shimojo, S. (2000). Illusions. What you see is what you hear. Nature *408*, 788.
20. Shams, L., Kamitani, Y., and Shimojo, S. (2002). Visual illusion induced by sound. Brain Res. Cogn. Brain Res. *14*, 147–152.
21. Berger, T.D., Martelli, M., and Pelli, D.G. (2003). Flicker flutter: is an illusory event as good as the real thing? J. Vis. *3*, 406–412.
22. Morein-Zamir, S., Soto-Faraco, S., and Kingstone, A. (2003). Auditory capture of vision: examining temporal ventriloquism. Brain Res. Cogn. Brain Res. *17*, 154–163.
23. Bertelson, P., Vroomen, J., de Gelder, B., and Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. Percept. Psychophys. *62*, 321–332.
24. Spence, C., and Driver, J. (2000). Attracting attention to the

illusory location of a sound: reflexive crossmodal orienting and ventriloquism. Neuroreport *11*, 2057–2061.
25. Jacobs, R.A., and Fine, I. (1999). Experience-dependent integration of texture and motion cues to depth. Vision Res. *39*, 4062–4075.
26. Triesch, J., Ballard, D.H., and Jacobs, R.A. (2002). Fast temporal dynamics of visual cue integration. Perception *31*, 421–434.
27. Efron, B., and Tibshirani, R.J. (1993). An Introduction to the Bootstrap (New York: Chapman & Hall).