Current Biology 22, 46–49, January 10, 2012 ©2012 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2011.11.039

Report

When Correlation Implies Causation in Multisensory Integration

Cesare V. Parise,[1,2,3,*] Charles Spence,[2] and Marc O. Ernst[1,3]
[1]Max Planck Institute for Biological Cybernetics and Bernstein Center for Computational Neuroscience, 72076 Tübingen, Germany
[2]Department of Experimental Psychology, University of Oxford, OX1 3UD Oxford, UK
[3]Cognitive Neuroscience Department and Cognitive Interaction Technology-Center of Excellence, Bielefeld University, 33615 Bielefeld, Germany

Summary

Inferring which signals have a common underlying cause, and hence should be integrated, represents a primary challenge for a perceptual system dealing with multiple sensory inputs [1–3]. This challenge is often referred to as the correspondence problem or causal inference. Previous research has demonstrated that spatiotemporal cues, along with prior knowledge, are exploited by the human brain to solve this problem [4–9]. Here we explore the role of correlation between the fine temporal structure of auditory and visual signals in causal inference. Specifically, we investigated whether correlated signals are inferred to originate from the same distal event and hence are integrated optimally [10]. In a localization task with visual, auditory, and combined audiovisual targets, the improvement in precision for combined relative to unimodal targets was statistically optimal only when audiovisual signals were correlated. This result demonstrates that humans use the similarity in the temporal structure of multisensory signals to solve the correspondence problem, hence inferring causation from correlation.

Results and Discussion

Multisensory signals originating from the same distal event are often similar in nature. Think of fireworks on New Year's Eve, an object falling and bouncing on the floor, or the footsteps of a person walking down the street. The temporal structure of the visual and auditory events are always almost overlapping, and we often effortlessly assume an underlying unity between our visual and auditory experiences. In fact, the similarity of temporal structure of unisensory signals provides a potentially powerful cue for the brain to solve the causal inference problem, i.e., to determine whether or not multiple sensory signals have a common origin.

One measure that is commonly used in signal processing to quantify the similarity between two signals as a function of their time lag is cross-correlation; the higher the cross-correlation between two signals at a given time lag, the higher their similarity. Cross-correlation (hereafter simply referred to as "correlation") is an important cue for humans when attempting to solve the correspondence problem in stereo vision [11], and a role has also been suggested in multisensory research [12].

*Correspondence: cesare.parise@tuebingen.mpg.de

Given that sensory integration should occur only if the underlying unisensory signals have a common cause, in the present study we generated trains of temporally correlated and uncorrelated audiovisual signals and assessed the strength of multisensory integration (MSI). If temporal correlation operates as a cue for causal inference, sensory fusion should occur only when the auditory and visual stimuli are temporally correlated.

Previous research has demonstrated that when multisensory stimuli are optimally integrated, the resulting percept ($\widehat{S}$) is a weighted average of the individual sensory estimates ($\widehat{S}_i$) with weights (w_i) proportional to their precision [13, 14]. If the noise of the individual sensory estimates (σ_i^2) is independent and Gaussian distributed, the maximum-likelihood estimate (MLE) of a physical property is:

$$\widehat{S} = \sum_i w_i \widehat{S}_i, \qquad \text{(Equation 1)}$$

where

$$w_i = \frac{1/\sigma_i^2}{\sum_j 1/\sigma_j^2}, \qquad \text{(Equation 2)}$$

and σ_i^2 is the variance of a sensory estimate $\widehat{S}_i$. Notably, if unimodal sensory cues are integrated according to MLE, the resulting sensory estimate should also be more precise than either of the individual sensory estimates, and its variance is given by

$$\sigma_{ij}^2 = \frac{\sigma_i^2 \sigma_j^2}{\sigma_i^2 + \sigma_j^2}. \qquad \text{(Equation 3)}$$

MLE therefore allows one to make clear predictions concerning the combined estimates, thereby providing a powerful benchmark against which to test for the effects of signal correlation on the optimality of multisensory integration.

To this end, ten observers were presented with streams of visual, auditory, and combined audiovisual stimuli that were sometimes correlated (Figure 1A, see Movie S1 available online). These stimuli, consisting of trains of white noise clicks (Aud.) and/or 30% contrast Gaussian blobs (Vis.), were presented from a variety of different spatial locations within a large 2D display. Observers were instructed to move a visual cursor controlled by a graphic tablet to the perceived location of the stimuli on the screen (Figure 1B). Note that in the bimodal trials, the position of the auditory and visual stimuli always physically coincided, even in the temporally uncorrelated trials, and participants were explicitly informed of this fact. If MSI is modulated by temporal correlation, one would expect observers to optimally integrate multisensory spatial cues only when the stimuli are correlated, not when they are uncorrelated. Given that participants were instructed to indicate the perceived location of the stimuli on a 2D display, this paradigm allowed us to test simultaneously for the effects of stimulus correlation on MSI in both the horizontal and vertical dimensions.

In accordance with previous studies, unimodal pointing responses revealed a uniform compression of space in vision and a vertical compression and horizontal expansion of space in audition [15, 16] (Figure 2A). Such a result is not surprising,

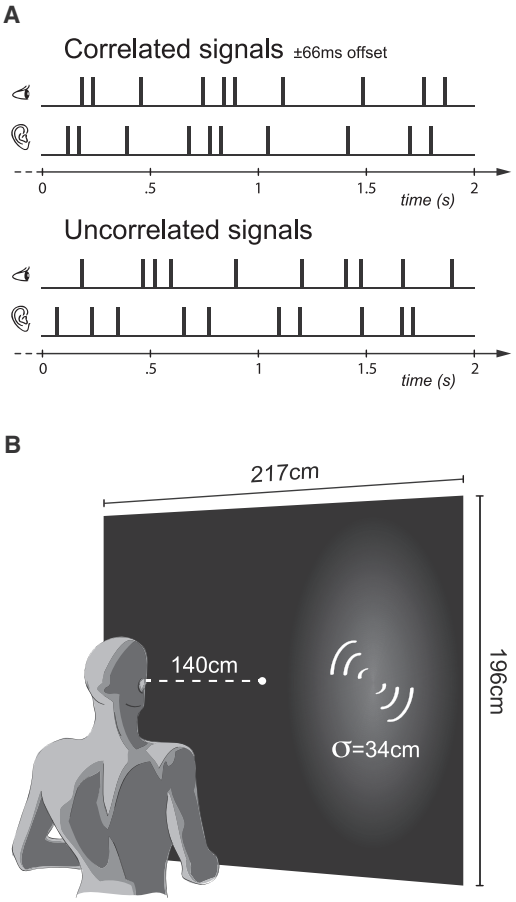

Figure 1. Stimuli and Apparatus

(A) Examples of the intensity profiles of correlated (correlation 1, lag = ±66 ms) and uncorrelated audiovisual stimulus pairs. Auditory stimuli consisted of trains of 10 white noise bursts. Visual stimuli consisted of trains of 10 Gaussian blobs. The overall duration of visual and auditory stimuli was 2 s.

(B) Schematic representation of the experimental apparatus.

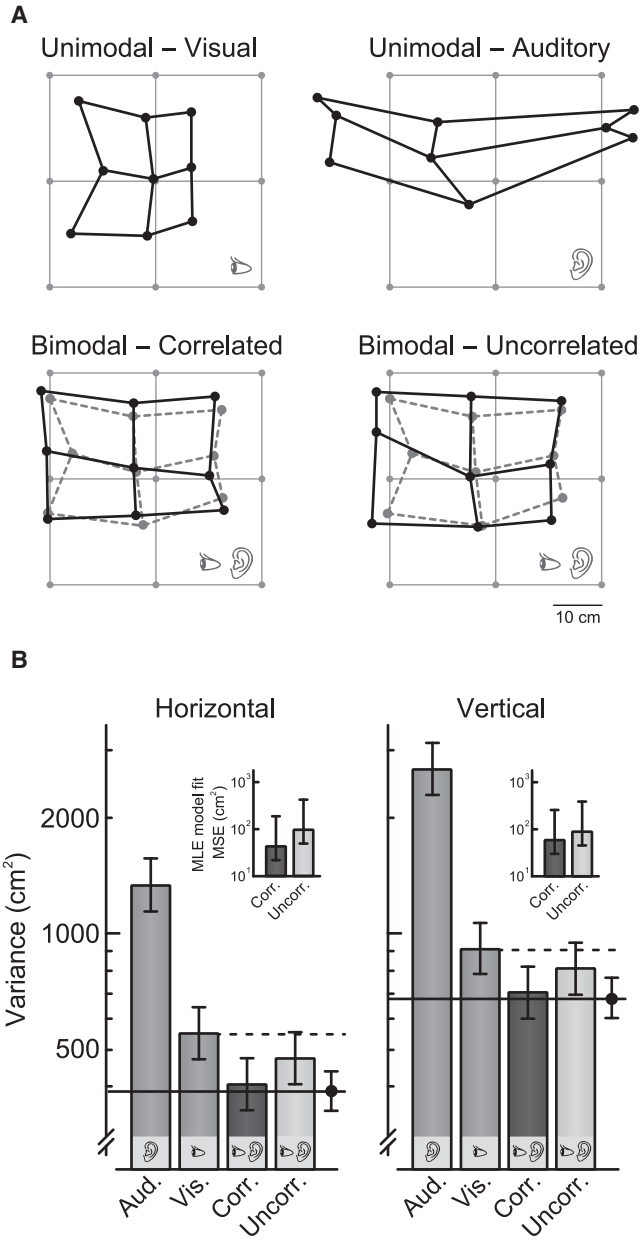

Figure 2. Results

(A) Average endpoint of pointing responses in the four conditions. The filled points correspond to the average end points. The black thin grid represents the actual position of the stimuli. In the lower two panels (bimodal condition), the dashed lines and gray dots represent the MLE prediction from Equation 1.

(B) Variance of the pointing responses in the four conditions in the horizontal (left) and vertical dimension (right). The dot on the right of each panel indicates the MLE prediction from Equation 3. The dashed line represents the variance of the most precise unimodal condition (i.e., vision). The inlaid bar plot represents the goodness of fit (mean squared error, MSE) of the MLE prediction (Equation 1) to the empirical average endpoint of pointing responses (see A, bottom two panels). Error bars indicate 95% confidence intervals.

given that it is well known that inaccuracies and distortions in perception can often be induced by prior knowledge [17] (and they are known to be especially pronounced under laboratory conditions such as when the head is fixed or when the localization cues are experimentally impoverished). As a consequence of these distortions, there was always a perceptual conflict between the visual and the auditory percept in the bimodal trials, even though no conflicts were physically present. For our purposes it does not matter how (in)accurate participants' unisensory percepts were; rather, the presence of "natural" conflicts allows us to investigate the weighting behavior as a result of integration.

To test for optimality, we first analyzed the precision of observers' localization responses to bimodal stimuli. In keeping with our hypothesis, they were more precise when the audiovisual stimuli were correlated than when they were not (Figure 2B). Crucially, the precision of participants' localization responses on those trials in which the visual and auditory stimuli were correlated were no different than predicted by MLE (Equation 3), whereas, on uncorrelated trials, precision was significantly lower than optimal in both the horizontal and vertical dimensions (see Supplemental Experimental Procedures). Yet, localization responses in the uncorrelated bimodal condition, though suboptimal, were still more precise than the best unimodal condition. This indicates some degree of residual multisensory integration. In this regard, it is important to note that here we manipulated the correlation between

the fine temporal structures of auditory and visual signals. However, when looking at the temporal structures of these stimuli on a coarser timescale (seconds), visual and auditory signals would nevertheless temporally coincide even in the uncorrelated condition. Moreover, given that we designed the uncorrelated stimuli by independently generating random temporal structures in vision and audition, sometimes also uncorrelated signals had nonzero correlation. It therefore seems reasonable to still find some degree of integration even in the uncorrelated condition.

The perceived position of the correlated stimuli was well predicted by the MLE weighting scheme (Equations 1, 2, and 3) in both the horizontal and vertical dimensions, as indicated by the gray grid in Figure 2A (lower row). This demonstrates the optimal integration of the correlated audiovisual stimuli. Interestingly though, the empirical visual weights in both the correlated $w_v(x, y) = (.62, .73)$ and uncorrelated $w_v(x, y) = (.66, .64)$ trials were very similar and in close agreement with optimal weights calculated from Equation 2, $w_v(x, y) = (.71, .74)$. Nevertheless, both the localization variance as well as the mean squared error (MSE) of the MLE model to the empirical data were higher for the uncorrelated than the correlated trials in both the vertical and horizontal directions (Figure 2, inset; see Supplemental Experimental Procedures). This may reflect the adoption of a switching strategy by the participants when the integrated percept started to break down with uncorrelated stimuli [18].

Given that the perceptual mechanisms for estimating the azimuth and the elevation of a sound source are distinct and based on different perceptual cues [19], and given that the pointing errors in the horizontal and vertical direction were uncorrelated (see Supplemental Experimental Procedures), the present data set allowed us to simultaneously make two independent assessments of the effect of stimulus correlation on MSI. Notably, the patterns of results emerging from the vertical and horizontal pointing errors were very similar, hence providing independent converging evidence for the effect under study.

These results demonstrate that human observers use the correlation between signals presented in different sensory modalities to integrate audiovisual cues. That is, whenever faced with auditory and visual signals whose fine temporal structures are correlated, the sensory system is more likely to infer a common underlying cause and to integrate the two sources of information optimally into a coherent representation of a single event. In other words, when crossmodal signals are correlated, observers seem to expect such stimuli to refer to the same distal event; conversely, when crossmodal signals are uncorrelated, they are more likely to be considered as being independent (that is, as belonging to different physical events).

It should be noted that these results cannot be explained in terms of audiovisual synchrony, with the auditory and visual stimulus streams being more asynchronous in the uncorrelated condition than in the correlated condition. To this end, we made sure to equate the correlated and uncorrelated conditions in terms of the average audiovisual delays by introducing a ±66 ms temporal offset (i.e., asynchrony) between the visual and auditory stimuli on each trial in the correlated condition (i.e., a cross-correlation with a maximum correlation = 1 at a lag of 66 ms). This offset value corresponds to the average temporal gap between the closest neighboring visual and auditory events in the uncorrelated bimodal signals. Therefore, the results of the present study demonstrate that it is the correlation between the temporal structures of the unisensory signals, rather than simply their (a)synchrony, that modulates audiovisual integration. It will be a challenge for future research to investigate the effect of different temporal offsets between correlated stimuli and the possible interactions between correlation and delay.

Our results demonstrate that MSI in the spatial dimension is influenced by the correlation between the signals along the (orthogonal) time dimension. In other words, the temporal correlation between multiple sensory signals promotes spatial MSI by informing the system that two signals have a common physical cause. Correlation provides a measure of statistical dependence between two variables: the higher the correlation, the more strongly two variables are related, but this by no means implies causation. Sensory systems, however, have no direct access to the causal structure of the real world, and therefore causality must be inferred from the available sensory cues. Therefore, knowing that two signals are correlated (i.e., not statistically independent), makes it more likely that the organism will assume a common underlying cause. In this sense, for the human sensory system, correlation really does imply causation.

Experimental Procedures

Nine naive observers and C.V.P. took part in the experiment. All of the participants had normal or corrected-to-normal vision and audition. The visual stimuli consisted of trains of flashes of large, low-contrast (30%) Gaussian blobs (sigma = 34 cm) back-projected against a black background on a matte plexiglass screen (size 217 × 196 cm). The auditory stimuli consisted of trains of white noise clicks delivered via earbuds. The participants were seated 140 cm from the screen, with their head constrained by a chinrest and a headrest and tightened by a temple-clamp.

Each train of visual and auditory stimuli consisted of ten flashes or ten clicks, respectively, randomly scattered over a 2 s temporal interval. Each click and blob had a duration of 16 ms. A new temporal structure (i.e., train of flashes and/or clicks) was generated for each trial. In the bimodal trials, where both visual and auditory stimuli were presented, the temporal structure was either identical in the two modalities (correlated trials) or random (uncorrelated trials, see Movie S1). Unimodal and bimodal trials were presented in separate blocks, preceded by an instruction screen informing the participants about the stimulus type. Auditory and visual trials alternated pseudorandomly in the unimodal blocks; correlated and uncorrelated audiovisual stimuli alternated pseudorandomly in the bimodal blocks of trials. Participants were not informed about the stimulus (un)correlation. To avoid the possibility of crossmodal temporal recalibration [20], visual-leading or auditory-leading trials pseudorandomly alternated in the bimodal block of trials. Each block consisted of 25 trials, and unimodal and bimodal blocks alternated during the experiment.

On each trial, the visual and auditory stimuli were presented pseudorandomly from 1 of 25 spatial positions arranged along a 5 × 5 two-dimensional grid (87.8 × 87.8 cm) aligned with observers' line of sight. The experimental trials consisted of the presentation of stimuli coming from one of the nine central positions in the grid. The external positions, closer to the edges of the screen, were included in order to broaden the stimulus space, thus increasing the positional uncertainty, and hence reduce any bias by participants to point straight ahead. Given that in the external position the visual stimuli were visibly clipped (i.e., the luminance profile of the outer stimuli was not Gaussian) and that the frame of the screen constrained participant responses, data from the outer position of the grid were not included in the analysis.

To create compelling spatialized sounds, in a preliminary session the raw auditory stimuli (clicks) were played from a loudspeaker from each of the 25 spatial positions and recorded with a pair of custom-built miniature microphones placed inside the left and the right ear canals of the blindfolded participants (see Figure 3). Tailored auditory stimuli were provided by performing this procedure individually for each participant. This ensured that the clicks were filtered by the individual's head-related transfer function (HRTF), thereby providing rich and ecological cues for sound localization. In order to have a large stimulus set and minimize the effect of potential

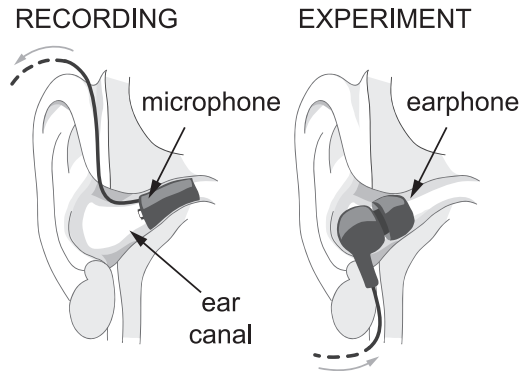

Figure 3. Microphone and Earphone Placement

During the recording of the stimuli, a pair of microphones was placed deep inside the ear canal of the observers in order to record the auditory stimuli physically played from each position of the stimulus grid. During the experiment, the earphones were placed inside the ear canal with the speaker positioned at the same depth as the microphones. Observers' head position was kept constant throughout the whole procedure.

artifacts in the recording procedure, eight clicks were recorded from each spatial position for each participant. On each trial, a new train of clicks was generated by randomly sampling with replacement from the set of clicks recorded from the relevant spatial position. The study was conducted in accordance with the Declaration of Helsinki and had ethical approval from the University of Tübingen.

Supplemental Information

Supplemental Information includes Supplemental Experimental Procedures and one movie and can be found with this article online at doi:10.1016/j.cub.2011.11.039.

Acknowledgments

We would like to thank V. Harrar for her help on a preliminary stage of this project. This study is part of the research program of the Bernstein Center for Computational Neuroscience, Tübingen, funded by the German Federal Ministry of Education and Research (BMBF; FKZ: 01GQ1002).

Received: September 12, 2011
Revised: October 18, 2011
Accepted: November 16, 2011
Published online: December 15, 2011

References

1. Welch, R.B., and Warren, D.H. (1980). Immediate perceptual response to intersensory discrepancy. Psychol. Bull. *88*, 638–667.
2. Ernst, M.O., and Bülthoff, H.H. (2004). Merging the senses into a robust percept. Trends Cogn. Sci. (Regul. Ed.) *8*, 162–169.
3. Shams, L., and Beierholm, U.R. (2010). Causal inference in perception. Trends Cogn. Sci. (Regul. Ed.) *14*, 425–432.
4. Bresciani, J.P., Ernst, M.O., Drewing, K., Bouyer, G., Maury, V., and Kheddar, A. (2005). Feeling what you hear: auditory signals can modulate tactile tap perception. Exp. Brain Res. *162*, 172–180.
5. Gepshtein, S., Burge, J., Ernst, M.O., and Banks, M.S. (2005). The combination of vision and touch depends on spatial proximity. J. Vis. *5*, 1013–1023.
6. Körding, K.P., Beierholm, U., Ma, W.J., Quartz, S., Tenenbaum, J.B., and Shams, L. (2007). Causal inference in multisensory perception. PLoS ONE *2*, e943.
7. Vatakis, A., and Spence, C. (2007). Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. Percept. Psychophys. *69*, 744–756.
8. Parise, C.V., and Spence, C. (2009). 'When birds of a feather flock together': synesthetic correspondences modulate audiovisual integration in non-synesthetes. PLoS ONE *4*, e5664.
9. Helbig, H.B., and Ernst, M.O. (2007). Knowledge about a common source can promote visual-haptic integration. Perception *36*, 1523–1533.
10. Radeau, M., and Bertelson, P. (1987). Auditory-visual interaction and the timing of inputs. Thomas (1941) revisited. Psychol. Res. *49*, 17–22.
11. Tyler, C.W. (1978). Binocular cross-correlation in time and space. Vision Res. *18*, 101–105.
12. Burr, D., Silva, O., Cicchini, G.M., Banks, M.S., and Morrone, M.C. (2009). Temporal mechanisms of multimodal binding. Proc. Biol. Sci. *276*, 1761–1769.
13. Ernst, M.O., and Banks, M.S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature *415*, 429–433.
14. Alais, D., and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. Curr. Biol. *14*, 257–262.
15. Grusser, O. (1983). Multimodal structure of the extrapersonal space. In Spatially Oriented Behavior, A. Hein and M. Jeannerod, eds. (New York: Springer-Verlag), pp. 327–352.
16. Lewald, J., and Ehrenstein, W.H. (1998). Auditory-visual spatial integration: a new psychophysical approach using laser pointing to acoustic targets. J. Acoust. Soc. Am. *104*, 1586–1597.
17. Knill, D.C., and Richards, W. (1996). Perception as Bayesian Inference (Cambridge, UK: Cambridge University Press).
18. Gepshtein, S., and Banks, M.S. (2003). Viewing geometry determines how vision and haptics combine in size perception. Curr. Biol. *13*, 483–488.
19. Blauert, J. (1997). Spatial Hearing: The Psychophysics of Human Sound Localization (Cambridge, MA: The MIT Press).
20. Fujisaki, W., Shimojo, S., Kashino, M., and Nishida, S. (2004). Recalibration of audiovisual simultaneity. Nat. Neurosci. *7*, 773–778.