Current Biology
Magazine

CellPress

Correspondence

Disambiguating vision with sound

Monica Gori[1], David Burr[1,2,3], and Claudio Campus[1]

An important task for the visual system is to identify and segregate objects from background. Figure–ground illusions, such as Edgar Rubin's bistable 'vase–faces illusion'[1], make the point clearly: we see either a central vase or lateral faces, alternating spontaneously, but never both images simultaneously. The border is perceptually assigned to either faces or vase, which become figure, the other shapeless background[2]. The stochastic alternation between figure and ground probably reflects mutual inhibitory processes that ensure a single perceptual outcome[3]. Which shape dominates perception depends on many factors, such as size, symmetry, convexity, enclosure, and so on, as well as attention and intention[4]. Here we show that the assignment of the visual border can be strongly influenced by auditory input, far more than is possible by voluntary intention.

Sixteen participants reported by continuous keypress their current perception (face or vase) of an image like Figure 1A. Four conditions were presented successively, each for 60 seconds, all repeated after a short break. The image was first presented with no sound, then accompanied by a conversation between two people, then by a dog barking, then saxophone music (soundtracks in Supplemental information). Figure 1B summarizes the results, averaged over sessions and participants. With no sound, participants saw faces or vase with equal probability (p(faces) = 0.51 ± 0.03, mean ± s.e.m.); while listening to the two conversing speakers, however, the proportion of time seeing faces increased to 0.78, a large and highly significant effect ($t(15) = 16.9$, $p < 10^{-7}$, $\log_{10}$ Bayes Factor > 8). The effect was equally strong for all 16 participants (individual data in Supplemental information). This was not a generic effect of concurrent auditory signals, as the dog and music soundtracks had no significant effect on figure predominance. Nor did the effect depend on presentation order (data not shown). Figure 1C shows the timecourse of the main experiment: as soon as the conversation started, the proportion of face percepts increased immediately to around 0.78, remaining constant until the soundtrack changed to barking dogs, when it plummeted back to near 0.5.

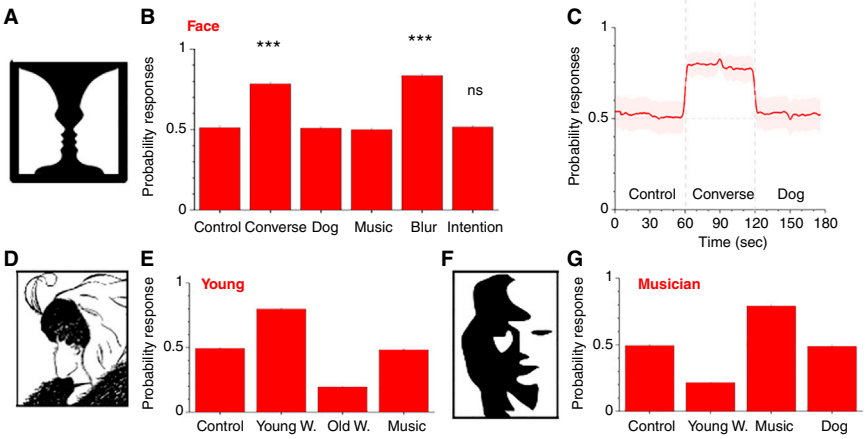

Figure 1. Auditory biasing of bistable visual figures.
(A) Ruben's vase-faces illusion appears to alternate perceptually between confronting faces and a central vase. (Vase/faces illusion inspired by 'Rubin's vase'.) (B) Probability of responding 'faces' for the six conditions described in the text: no soundtrack; listening to conversation; dog barking; saxophone music; conversation with the image blurred; intention to see faces. Bars show ±1 s.e.m. (C) Timecourse of the first three conditions of the main experiment (see Supplemental Information for others). Averaged data are smoothed by gaussian window of 5 sec time constant. Shaded area ±1 s.e.m. (D) Edward Boring's 'old–young woman' illusion. (E) Average proportion of 'young' reports for the four conditions described in the text: no soundtrack; listening young woman talking; old woman talking; saxophone music. (F) 'Sax player–face illusion'. (Adapted from 'Sara Nader' by Roger Shepard.) (G) Average proportion of reports of 'sax player' for the four conditions described in the text: no soundtrack; listening to young woman talking; saxophone music; dog barking.

We then measured the same participants with the image blurred to reduce the reliability of visual information. With blurred stimuli the effect of speech on face perception was even greater, increasing face domination to 0.84, significantly more than for unblurred stimuli ($t(15) = 5.9$, $p < 0.001$, $\log_{10}$BF > 2). The increase in the auditory effect is predicted by most ideal-observer models of multi-sensory perception[5,6], as discussed below.

We tested whether the effects could be ascribed to attention or intention[4], rather than to genuine audio-visual integration. In a separate session with no sound, participants were instructed to attend to and 'try to see' the faces rather than the vase in the second minute of the experiment, then to do the converse. The instructions produced had no significant effect on the perceptual outcome (ANOVA for three conditions: $F(2,22) = 1.42$, $p = 0.26$, $\log_{10}$BF = −0.18). Previous research has reported significant effects of intention and attention on bistability, but they are typically weak[4] .

We replicated the results with two other well-known bistable illusions: Boring's 'young–old woman illusion' and the 'sax player–face illusion'. Both illusions were highly influenced by hearing a concurrent, relevant sound. The young–old woman illusion biased heavily to 'young' with the voice of a young woman and to 'old' with the voice of an old woman, while music had no effect. The sax player–face illusion was also appropriately modulated by the young woman's voice and the sax music, but not by the dog barking. Both illusions were stronger when the stimuli were blurred than when sharp (see Supplemental information).

These results provide strong evidence for the role of multi-modal integration in figure–ground border assignment, an essential precursor to object segregation. The illusions shown here were deliberately designed to maximize ambiguity, but real-world images can also be noisy and ambiguous. Under difficult conditions, multi-sensory information can be invaluable, as different senses access complementary sources of information, which can vary with conditions (for example, sound can be more useful at night). There is now general consensus that the brain optimizes perception by combining information from different senses, weighting each by its reliability[5,6]. This theory predicts that if the visual information is degraded, making it less reliable, auditory input should be weighted more highly. We tested the idea by blurring the face–vase image: as predicted, the auditory input became significantly more effective.

That strong biases were induced by appropriate auditory soundtracks, but not by intentional effort, suggests that auditory input acts directly on the visual information, rather than indirectly via attention or cognitive bias. If the auditory effects were mediated simply by evoking a cognitive bias towards faces, they should produce similar results to the intentional condition, where observers were explicitly directed to do just that, to try to see faces rather than the vase: yet the instructional effect was minimal and insignificant. That auditory signals, but not conscious volition, can profoundly affect object visual border assignment and object segregation suggests an early interaction, at sensory rather than decisional levels. This is consistent with previous evidence of multisensory interactions with binocular rivalry (another example of bistable visual perception), where touching haptic gratings biased the rivalry in favour of the orientation of the touched grating[7]. The bias was selective for grating spatial frequency, suggesting early sensory interaction, probably in primary visual cortex (area V1).

There are countless examples of compelling multi-sensory interactions, most notably the McGurk effect[8], where observing lip movements affects what is heard, and the ventriloquist effect[5], where a sound appears to emanate from the dummy rather than from the ventriloquist. But the current result is different, in that the auditory stimulus is not a simple tone or phoneme, but speech or music to be understood semantically. The sound source needs to be perceived and classified as a conversation between two speakers to bias the vase–faces percept, and associated with the speaker's age for the young–old illusion. However, although semantic coding is essential, the soundtrack does not bias the visual percept via cognitive processes, but seems to act directly on the visual information. Perhaps listening to the soundtracks creates — or directly reinforces — a multisensory neural representation related to the auditory content, and this happens only while there is actual sensory input.

The mechanisms for this suggested audio-visual interaction are far from clear, but there is much evidence for auditory signals activating visual areas[9]. Complex auditory input can activate visual areas, such as the visual word form area in blind participants trained to interpret auditory landscapes of visual forms[10]. Whatever the neural mechanisms underlying these effects, this study shows there is considerable interaction between the auditory and visual signals, serving to create a clear, unambiguous perceptual representation of the world in the face of noise and uncertainty.

SUPPLEMENTAL INFORMATION

Supplemental information includes one figure, one table, experimental procedures, and five audio files, and can be found with this article online at https://doi.org/10.1016/j.cub.2024.01.043.

A video abstract is available at http://doi.org/10.1016/j.cub.2024.01.043.#mmc7.

ACKNOWLEDGEMENTS

Supported by ERC-Stg 948349 "MySpace" to M.G. and ERC-Adv grant 832813 "GenPercept" to D.B.

AUTHOR CONTRIBUTIONS

M.G. conceived the basic experiment. M.G. and C.C. collected the data. All authors participated in designing the experiment, analysing the data and writing the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Rubin, E. (1921). Visuell wahrgenommene figuren: Studien in psychologischer analyse (Copenhagen: Gyldendalske boghandel).
2. Peterson, M., and Salvagio, E. (2010). Figure-ground perception. Scholarpedia *5*, 4320.
3. Peterson, M.A., and Skow, E. (2008). Inhibitory competition between shape properties in figure-ground perception. J. Exp. Psychol. Hum. Percept. Perform. *34*, 251–267.
4. Suzuki, S., and Peterson, M.A. (2000). Multiplicative effects of intention on the perception of bistable apparent motion. Psych. Sci. *11*, 202–209.
5. Alais, D., and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. Curr. Biol. *14*, 257–262.
6. Trommershauser, J., Kording, K., and Landy, M.S. (2011). Sensory Cue Integration (Oxford: Oxford University Press).
7. Lunghi, C., Binda, P., and Morrone, M.C. (2010). Touch disambiguates rivalrous perception at early stages of visual analysis. Curr. Biol. *20*, R143–R144.
8. McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. Nature *264*, 746–748.
9. Gori, M., Bertonati, G., Campus, C., and Amadeo, M.B. (2023). Multisensory representations of space and time in sensory cortices. Hum. Brain Mapp. *44*, 656–667.
10. Striem-Amit, E., Cohen, L., Dehaene, S., and Amedi, A. (2012). Reading with sounds: Sensory substitution selectively activates the visual word form area in the blind. Neuron *76*, 640–652.

[1]UVIP — Unit for visually impaired people, Italian Institute of Technology, Genoa 16152, Italy. [2]Department of Neuroscience, University of Florence, Florence 50135, Italy. [3]School of Psychology, University of Sydney, Camperdown, NSW 2050, Australia.
E-mail: monica.gori@iit.it (M.G.); davidcharles.burr@unifi.it (D.B.); claudio.campus@iit.it (C.C.)