Conscious access to the unisensory components of a cross-modal illusion

Salvador Soto-Faraco[a,b] and Agnès Alsius[b,c]

[a]ICREA, [b]Parc Científic de Barcelona and [c]Departament de Psicologia Básica, University of Barcelona, Spain

Correspondence to Dr Salvador Soto-Faraco, Parc Científic de Barcelona, Hospital Sant Joan de Déu (Edifici Docent), c/Santa Rosa, 39-57, Planta 4ª, 08950 Esplugues de Llobregat (Barcelona), Spain
Tel: +34 93 6009769; fax: +34 93 6009768; e-mail: Salvador.Soto@Icrea.Es

Received 6 November 2006; accepted 26 November 2006

It is often claimed that binding information across sensory modalities leads to coherent, unitary mental representations. The dramatic illusions experienced as a result of intersensory conflict, such as the McGurk effect, are often attributed to a propensity of the perceptual system to impose multisensory coherence onto events originating from a common source. In contrast with this ssumption of unity, we report an unexpected ability to resolve the timing between sound and sight regarding multisensory events that induce an illusory reversal of the elements specified in each modality. This finding reveals that the brain can gain access, simultaneously, to unisensory component information as well as to the result of the integrated multisensory percept, suggesting some degree of penetrability in the processes leading to cross-modality binding. *NeuroReport* 18:347–350 © 2007 Lippincott Williams & Wilkins.

Keywords: attention, cognitive neuroscience, humans, multisensory integration, speech

Introduction

Information processing at early stages of the nervous system involves segregated, sensory specific, pathways (e.g. law of specific energies [1]). Yet, as the objects of perception are characterized by features in several sensory modalities, binding information across the senses becomes critical to represent the world [2]. Multisensory binding processes can take place quite early in functional terms. For example, electrophysiological correlates of multisensory integration in humans can be detected just 40 ms after stimulus onset [3,4]. Similarly, blood oxygenation level-dependent (BOLD) responses measured in the human brain using functional MRI reveal cross-modal activity in areas traditionally considered unisensory [5], a finding which is consistent with (though it does not necessarily imply) effects at early processing stages. Converging with these neural correlates, psychophysical findings suggest that multisensory integration processes can occur automatically, without much degree of voluntary control [6–9]. This concurs with claims that integration mechanisms are rapid and mandatory, so the resulting representations can be made quickly available to neural control systems for decision making or motor response.

A classic illustration of multisensory binding relates to the intimate link between seen and heard aspects of communication. Indeed, seeing articulatory gestures of the speaker improves oral comprehension of spoken messages [10,11]. At the neural level, audiovisual speech leads to increased BOLD responses in the heteromodal cortex (e.g. superior temporal sulcus [12]), and to modulation of early auditory-evoked electrophysiological potentials (N1/P2 complex [13,14]). The classic behavioral demonstration of audio-visual speech binding is the McGurk illusion [15]; when a sound (e.g. /ba/) is dubbed onto mismatching lip-movements (e.g. [ga]) the observer hears the visually specified syllable, or even a mixture (fusion) between sound and sight (in this case *da*). Proposals for a pre-attentive and automatic view of multisensory integration [16] have often argued that these interactions operate in a modular [17] and cognitively impenetrable [18] fashion. According to this view, there should be no, or very limited, access to information about the original sensory components once cross-modal integration has occurred [8,9]. Here, we challenge this hypothesis by showing temporal order sensitivity between heard and seen speech in situations when multisensory binding leads to an illusory percept.

We used videoclips in which the sound of the syllable /da/ was dubbed onto the lip gestures of [ba], a combination that often leads observers to hear the sequence *bda* (as in ab*duc*tion). The alternate sequence *dba* (as in od*dba*ll) is seldom experienced, arguably because the shape of the lips while uttering [ba] is inconsistent with the sound /dba/. The stimulus onset asynchrony (SOA) between audio and video was varied to evaluate how much the acoustic component /da/ could lead the visual component [ba], before the expected *bda* percept would break down. Participants were asked to report what they heard and to judge the temporal order regarding the visual and the acoustic signals [so we could find the time interval needed to resolve the order of modalities, the just noticeable difference (JND)]. According to the automaticity view, multisensory integration will impose a limit on temporal order resolution between unisensory events, which should abide to the sequence experienced as a result of the illusion.

If we find SOAs in which participants correctly report that audition came first (i.e. /da/ before [ba]) but fall to the illusion sequence (*bda*), then the encapsulated, cognitively impenetrable, view of multisensory binding will need to be qualified.

Methods

We tested 50 participants in three experiments after informed consent (n=18, 21, and 10, respectively). They were presented with dubbed videoclips (720 × 576 pixels; showing the lower half of the speaker's face) of 3 s (75 frames) duration. The first facial movement always occurred at the 22nd frame (840 ms). The sound of /ba/ lasted for 390 ms and /da/ for 320 ms. In experimental trials, the sound of /da/ was dubbed onto the visual [ba] (vice-versa in Experiment 3); in filler trials (not analyzed), visual and acoustic components coincided (both 'da' or 'ba'). White noise (65/50 dB signal/noise ratio) was added to the audio track. Audiovisual synchronization was achieved by aligning in time the consonant bursts of the two original audio signals (this was SOA=0). Each clip was produced by desynchronizing the visual and audio channels by 13 different SOAs: −640, −400, −320, −240, −160, −80, 0, 160, 240, 320, 400, 480, and 720 (negative indicates audio-lead). Experiment 1 contained 195 trials, whereas Experiments 2 and 3 contained 260 trials, presented in runs of 65 trials randomly ordered (three experimental plus two fillers per each SOA).

A 22′ CRT monitor (100 Hz) and two loudspeakers (left and right of the monitor) were used to present the videoclips. After each clip participants were prompted for response. In the identification task, participants chose the heard syllable from five alternatives (*bda*, *da*, *ba*, *dba*, and *other*). In the temporal order task, participants judged the order of video and sound channels (leading or lagging channel, instruction counterbalanced). In Experiment 1, each task was performed in separate blocks (order counterbalanced), whereas in Experiments 2 and 3 both tasks were performed in each trial (order counterbalanced).

Results

In Experiment 1 (Fig. 1a), the dominant percept at the extreme SOAs (−640 and +720 ms) was *da*, the acoustically specified event [the contrast prop.(*da*) > prop.(*bda*) was $P = 0.002$ and 0.086, respectively]. As expected, when the auditory /da/ and visual [ba] events were presented around synchrony, visually dominated illusions (*bda* or *ba*) were common: in particular, *bda* percepts peaked at +160 ms (vision leading), and were equally or more frequent than *da* percepts from −400 to +480 ms [contrast prop.(*da*) > prop.(*bda*) at $P > 0.25$]. For the temporal order judgement (TOJ) analysis, we fitted a logistic curve to the proportion of visual-first responses across SOAs (all participants, $r^2 > 0.65$). The point of subjective equality (PSE; SOA with prob.=0.5) was achieved when vision led audition by +57 ms (SD = 75), a typical result in audiovisual TOJ. The just noticeable difference [JND (time interval yielding 75% accuracy)] was 152 ms (SD = 36), with the 75% performance limits at −94 and +208 ms, a common asymmetry in audiovisual speech [20]. Remarkably, audio leads larger than 94 ms allowed observers to judge reliably that sound (containing /da/) was ahead of video (containing [ba]) and yet, they often reported hearing *bda* (i.e. the visually specified phoneme before the acoustically specified one). In the SOA range from −400 to −160 ms, the probability of audio-first TOJ responses was 0.90, but the percept *bda* was reported on 30% of trials (an extra 12% responses were *ba*, the other visually dominant percept).

These results reveal an unexpected resolution in cross-modal temporal order despite audiovisual integration leading to illusory percepts inverting the order of acoustically and visually specified phonemes. This 'dual perception' account could, however, be negated on the basis of strategic

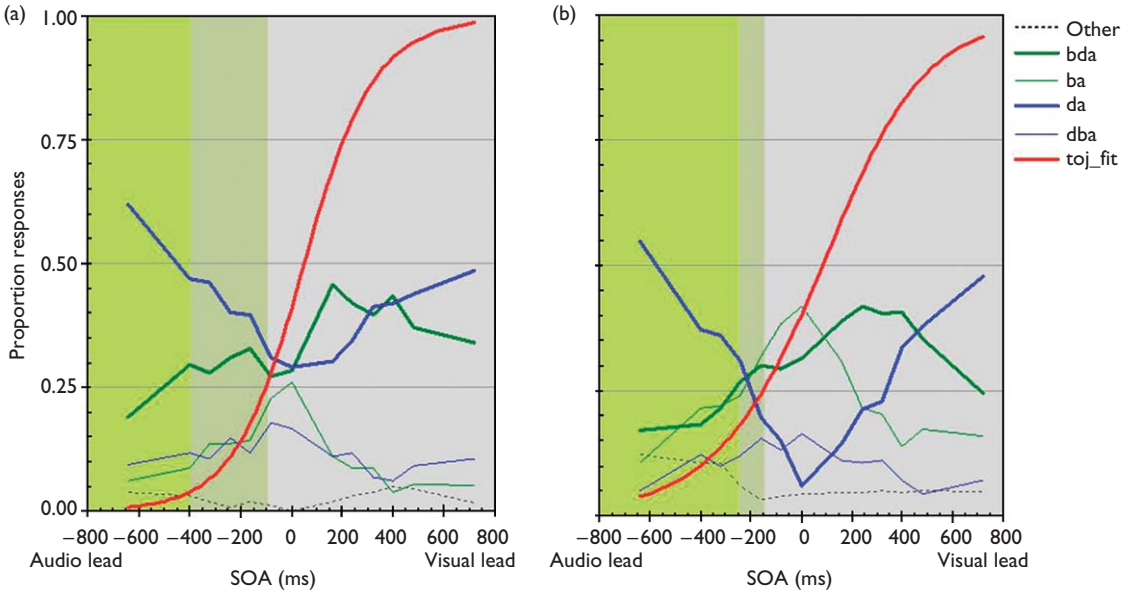

Fig. 1 Results of Experiments 1 (graph a) and 2 (graph b), representing the proportion of identification responses of each category (*bda*, *da*, *dba*, *ba*, and *other*, see color key) and logistic fit of the mean proportion of visual-first responses in the TOJ task (solid red line). The SOA range in which participants systematically responded that audition came first is shaded in green. The middle (darker) green area corresponds to the range of SOAs in which, despite systematically responding that sound came before video, the visually dominated response *bda* was as frequent as *da*, or more.

responding. In identification blocks, participants may engage in a speech perception mode that would promote integration [21], whereas in TOJ blocks they might try and segregate the sensory channels, thereby reducing the chances of illusory percepts. To test this alternative, we conducted a new experiment in which participants performed both the identification and TOJ tasks on each trial (reports about syllable identity and modality order were based on the same event).

In Experiment 2 (Fig. 1b), identification data revealed that at the extreme SOAs participants predominantly heard *da*, the acoustically specified event [contrasts prop.(*da*) > prop.(*bda*), $P<0.05$]. The illusory percept *bda* peaked at +240 ms, and it was as frequent as (or more frequent than) *da* from −240 to +480 ms [contrasts prop.(*da*) > prop.(*bda*) at $P>0.25$]. In the TOJ task, PSE was +74 ms (visual lead) and JND was 219 ms (limits at −144 and 293 ms). The larger JND seen here, with respect to Experiment 1, can be attributed to a cost of dual task. The critical finding, however, was that participants could correctly order the acoustic channel (/da/) before the visual channel ([ba]), and yet they frequently reported hearing the illusory sequence *bda*. This occurred at SOAs from −240 to −160 ms in which, despite high audio-first responses in the TOJ task (0.79), visually dominated identification responses prevailed, at 56% (*bda*=28%; *ba*=28%; acoustically dominant responses *da* and *dba* were at 25 and 14%, respectively).

In Experiment 2, both judgments were based on the very same event, and still observers had access to precise temporal information regarding individual modalities (responding that sound, /da/, led vision [ba]) as well as to the percept resulting from their binding (hearing, *bda*). If this striking phenomenon of dual perception is true, then such a paradoxical behavior should occur even at single trial level, not only on the average data. We conditioned the identification analysis to only those trials in which participants responded that audition came first in the TOJ task, at the critical asynchronies (−240 and −160 ms, in which the proportion of audition-first responses was 0.79). Among these trials, the proportion of visuallydominated identification responses was 59% (*bda*=30%; *ba*, the other visually induced percept was 29%; *da* and *dba* percepts achieved 25 and 12%, respectively). Thus, experiencing an illusory cross-modal percept that involves a reversed temporal sequence does not seem to constrain judgments on the physical order of the sensory events.

The use of combination illusions (/da/ + [ba]=*bda*) allowed us to expose the temporal sequence of the illusory percept (*bda*; different from *dba*), a critical aspect of the argument in Experiments 1 and 2. One might argue, however, that combinations are a phenomenologically weak instance of audiovisual speech binding, when compared with fusions (/ba/ + [ga]=*da*). In Experiment 3, the sound /ba/ was dubbed with the video [ga] to create fusion illusions (experimental trials). In the identification task, the extreme SOAs resulted in acoustic dominance [prop.(*ba*)=0.74 and 0.83, different from prop.(*da*) at $P<0.001$]. The illusory percept *da* peaked at 0 ms, and it was as frequent as, or more frequent than, the acoustic percept *ba* from −160 to +320 ms. In the TOJ task, PSE was +114 ms, and JND=234 ms (limits at −120 and 347 ms). As in Experiments 1 and 2, there were SOAs (around 160 ms) in which a high proportion of illusory percepts (*da*=0.33; equivalent to audio *ba* responses, $P>0.25$) coexisted with correct temporal order judgments ($P>0.75$).

Discussion

We report a coexistence between access to unisensory events (their temporal order can be determined) and multisensory binding between them (integration leads to an illusion). This finding questions the theory that the binding of sensory information takes place in a modular fashion whereby only the percept arising from integration becomes accessible to awareness [6–9,15]. Robust evidence exists for early expression of multisensory integration in terms of neurophysiology [3,4,19] and behavior [3,8]. Thus, although not inconsistent with an automatic view of multisensory integration, the present data suggest a flexible conception of these binding processes, allowing for some top-down influences. In particular, the encapsulated and cognitively impenetrable conceptualization of cross-modal processes, leading to the widespread assumption that multisensory binding negates access to unimodal information, needs to be revised.

Recent evidence favors the malleable nature of sensory binding. First, brain areas responsive to cross-modal asynchrony detection [22] do not overlap with traditional loci of cross-modal integration (see [12]). Indeed, a recent functional MRI study [23] suggests a dissociation between brain networks involved in processing cross-modal coincidence of physical events and those involved in cross-modality fusion. Second, our results fit well with recent suggestions that attention plays a modulatory role in mechanisms of cross-modal binding, challenging the idea that multisensory integration is pre-attentive [4,24]. Alsius *et al.* [24] showed that the McGurk illusion decreases under conditions of high attention load. Talsma and Woldorff [4] reported weaker electrophysiological correlates of multisensory integration when beep-flash pairs appeared at unattended, as opposed to attended, regions of space. Reconciling findings of early binding with the present demonstration would involve accepting that multisensory integration results, in part, from bidirectional interactions between higher order association brain regions and early sensory areas [5,12]. It is revealing that neurons in the cat superior colliculus, a neurophysiological model of cross-modal integration, display integrative response properties only if projections from the cortex are functional [25].

Conclusion

Our findings support the notion that multisensory integration processes are not fully automatic and opaque to awareness, but malleable through top-down processes. This is, in turn, consistent with proposals that attention plays a role during cross-modal binding and that the functional architecture of brain mechanisms subserving multisensory integration is heavily recurrent.

Acknowledgements

This research was supported by the *Ministerio de Educación y Ciencia* (Spain) TIN2004-04363-C03-02. A.A. is supported by a BRD scholarship from the University of Barcelona.

References

1. Müller J. *Handbuch der Physiologie des Menschen für Vorlesungen* [In London]. Coblenz, Holscher; translated by William Baly as elements of physiology 1840; vol. 2.
2. Stein BE, Meredith MA. *The merging of the senses*. Cambridge, Massachusetts: MIT Press; 1993.

3. Molholm S, Ritter W, Murray MM, Javitt DC, Schroeder CE, Foxe JJ. Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Brain Res Cogn Brain Res* 2002; **14**:115–128.
4. Talsma D, Woldorff MG. Selective attention and multisensory integration: multiple phases of effects on the evoked brain activity. *J Cogn Neurosci* 2005; **17**:1098–1014.
5. Calvert GA, Bullmore ET, Brammer MJ, Campbell R, Williams SC, McGuire PK, *et al*. Activation of auditory cortex during silent lipreading. *Science* 1997; **276**:593–596.
6. Bertelson P, Vroomen J, de Gelder B, Driver J. The ventriloquist effect does not depend on the direction of deliberate visual attention. *Percept Psychophys* 2000; **62**:321–332.
7. Driver J. Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature* 1996; **381**:66–68.
8. Soto-Faraco S, Navarra J, Alsius A. Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. *Cognition* 2004; **92**:B13–B23.
9. Soto-Faraco S, Spence C, Kingstone A. Assessing automaticity in the audiovisual integration of motion. *Acta Psychol* 2005; **118**:71–92.
10. Sumby W, Pollack I. Visual contribution to speech intelligibility in noise. *J Acoust Soc Am* 1954; **26**:212–215.
11. Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, Foxe JJ. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cereb Cortex* 2006; doi:10.1093/cercor/bhl024.
12. Calvert GA. Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cereb Cortex* 2001; **11**:1110–1123.
13. Besle J, Fort A, Delpuech C, Giard MH. Bimodal speech: early suppressive visual effects in human auditory cortex. *Eur J Neurosci* 2004; **20**: 2225–2234.
14. Klucharev V, Möttönen R, Sams M. Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Brain Res Cogn Brain Res* 2003; **18**:65–75.
15. McGurk H, MacDonald J. Hearing lips and seeing voices. *Nature* 1976; **265**:746–748.
16. Bertelson P, deGelder B. The psychology of multimodal perception. In: Spence C, Driver J, editors. *Crossmodal space and crossmodal attention*. Oxford: Oxford University Press; 2004. pp. 141–179.
17. Fodor J. *The modularity of mind*. MIT Press; 1983.
18. Pylyshyn Z. Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behav Brain Sci* 1999; **22**:341–365.
19. Sams M, Aulanko R, Hämäläinen M, Hari R, Lounasmaa OV, Lu ST, Simola J. Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neurosci Lett* 1991; **127**:141–145.
20. Munhall KG, Gribble P, Sacco L, Ward M. Temporal constraints on the McGurk effect. *Percept Psychophys* 1996; **58**:351–362.
21. Tuomainen J, Andersen TA, Tiippana K, Sams M. Audio-visual speech perception is special. *Cognition* 2005; **96**:B13–B22.
22. Bushara KO, Grafman J, Hallett M. Neural correlates of auditory-visual stimulus onset asynchrony detection. *J Neurosci* 2001; **21**:300–304.
23. Miller LM, D'Esposito M. Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *J Neurosci* 2005; **25**:5884–5893.
24. Alsius A, Navarra J, Campbell R, Soto-Faraco S. Audiovisual integration of speech falters under high attention demands. *Curr Biol* 2005; **15**:839–843.
25. Stein BE, Wallace MT, Vaughan JW, Jiang W. Crossmodal spatial interactions in subcortical and cortical circuits. In: Spence C, Driver J, editors. *Crossmodal space and crossmodal attention*. Oxford: Oxford University Press; 2004.