Current Biology, Vol. 15, 839–843, May 10, 2005, ©2005 Elsevier Ltd All rights reserved. DOI 10.1016/j.cub.2005.03.046

Audiovisual Integration of Speech Falters under High Attention Demands

Agnès Alsius,[1] Jordi Navarra,[1] Ruth Campbell,[2] and Salvador Soto-Faraco[1,*]
[1]Cognitive Neuroscience Group
Parc Científic de Barcelona
Departament de Psicologia Bàsica
Universitat de Barcelona
Passeig de la Vall d'Hebron, 171
08035 Barcelona
Spain
[2]Department of Human Communication Sciences
University College London
Chandler House
2 Wakefield Street
London WC1 N 1 PF
United Kingdom

Summary

One of the most commonly cited examples of human multisensory integration occurs during exposure to natural speech, when the vocal and the visual aspects of the signal are integrated in a unitary percept. Audiovisual association of facial gestures and vocal sounds has been demonstrated in nonhuman primates [1] and in prelinguistic children [2], arguing for a general basis for this capacity. One critical question, however, concerns the role of attention in such multisensory integration. Although both behavioral and neurophysiological studies have converged on a preattentive conceptualization of audiovisual speech integration [3–8], this mechanism has rarely been measured under conditions of high attentional load, when the observers' attention resources are depleted [9]. We tested the extent to which audiovisual integration was modulated by the amount of available attentional resources by measuring the observers' susceptibility to the classic McGurk illusion [3] in a dual-task paradigm [10]. The proportion of visually influenced responses was severely, and selectively, reduced if participants were concurrently performing an unrelated visual or auditory task. In contrast with the assumption that crossmodal speech integration is automatic, our results suggest that these multisensory binding processes are subject to attentional demands.

Results and Discussion

Participants were presented with a videotape of a female talker who occasionally pronounced a word in which the video and auditory channels had been dubbed in order to produce the McGurk illusion (see the Supplemental Data available with this article online). In this illusion, exposure to mismatched auditory and visual speech (lip movements) signals can lead observers to experience ("hear") a word reflecting the visual, rather than auditory, properties of the speech item or a "fusion," which incorporates some acoustic and some visual phonetic properties of the observed speech act (see Figures 1A and 1B). All participants were asked to repeat back what the speaker said under three different display conditions: audiovisual, visual alone, or auditory alone (see Figure 1C). The amount of available attentional resources was manipulated by a concurrent task, performed by half the participants (dual-task group). In Experiment 1, the concurrent task was performed on a visual stream, whereas in Experiment 2, the concurrent task was auditory. The remaining participants were shown the same displays but asked to simply view the monitor and repeat back the words (single-task group). The results for the concurrent repetition-detection task did not reveal any differential performance across experiments or conditions (see Supplemental Data). Regarding word recall, participants performed very accurately overall in the auditory and audiovisual conditions—they made predominantly the expected auditory or visual response—whereas they performed poorly in the visual-only condition (see Table 1).

The dependent variable of interest was the proportion of visual or fusion responses (i.e., illusory McGurk responses) given by the participants as a function of Display Condition (audiovisual, auditory, or visual) and Task (single or dual). The data were submitted to two mixed analyses of variance (ANOVA; significance levels were Greenhouse-Geisser corrected when appropriate), one with participants as the random factor, and the other with items as the random factor (denoted by subindexes 1 and 2, respectively).

In Experiment 1 (see Figure 2 for the group averages), the critical interaction between Task and Display Condition was statistically significant ($F_1 = 11.5$, $p = 0.001$; $F_2 = 31.1$, $p < 0.001$). In the audiovisual condition, the percentage of participants' visual/fusion responses was *significantly reduced* in the dual-task group (8.5%) as compared to the single-task group (33%; $t_1 = 3.8$, $p < 0.005$; $t_2 = 6.4$, $p < 0.001$). This result suggests that when participants focused attention on a difficult visual task, even when directly looking at the speaker's face, McGurk illusions were almost eliminated (indeed, in the dual-task group, the percentage of visual responses in the audiovisual conditions was equivalent to the auditory alone conditions; $|t| < 1$). In contrast, no differences as a function of task were found in either the visual-only ($t_1 = 0.9$, $p = 0.379$; $t_2 = 1.3$, $p = 0.181$) or auditory-only conditions ($t_1 = 0.5$, $p = 0.625$; $t_2 = 0.7$, $p = 0.477$). The visual-only condition produced hardly any correct responses, confirming that information available from silent speech actions alone is usually insufficient to enable word identification under open-response set conditions [11, 12]. The auditory-only condition did not show any effect of dual task, even when considering the proportion of auditory-based responses (see Table 1). The high accuracy level in the auditory control condition not only shows that words could be correctly identified at a perceptual level, but importantly that per-

*Correspondence: ssoto@ub.edu

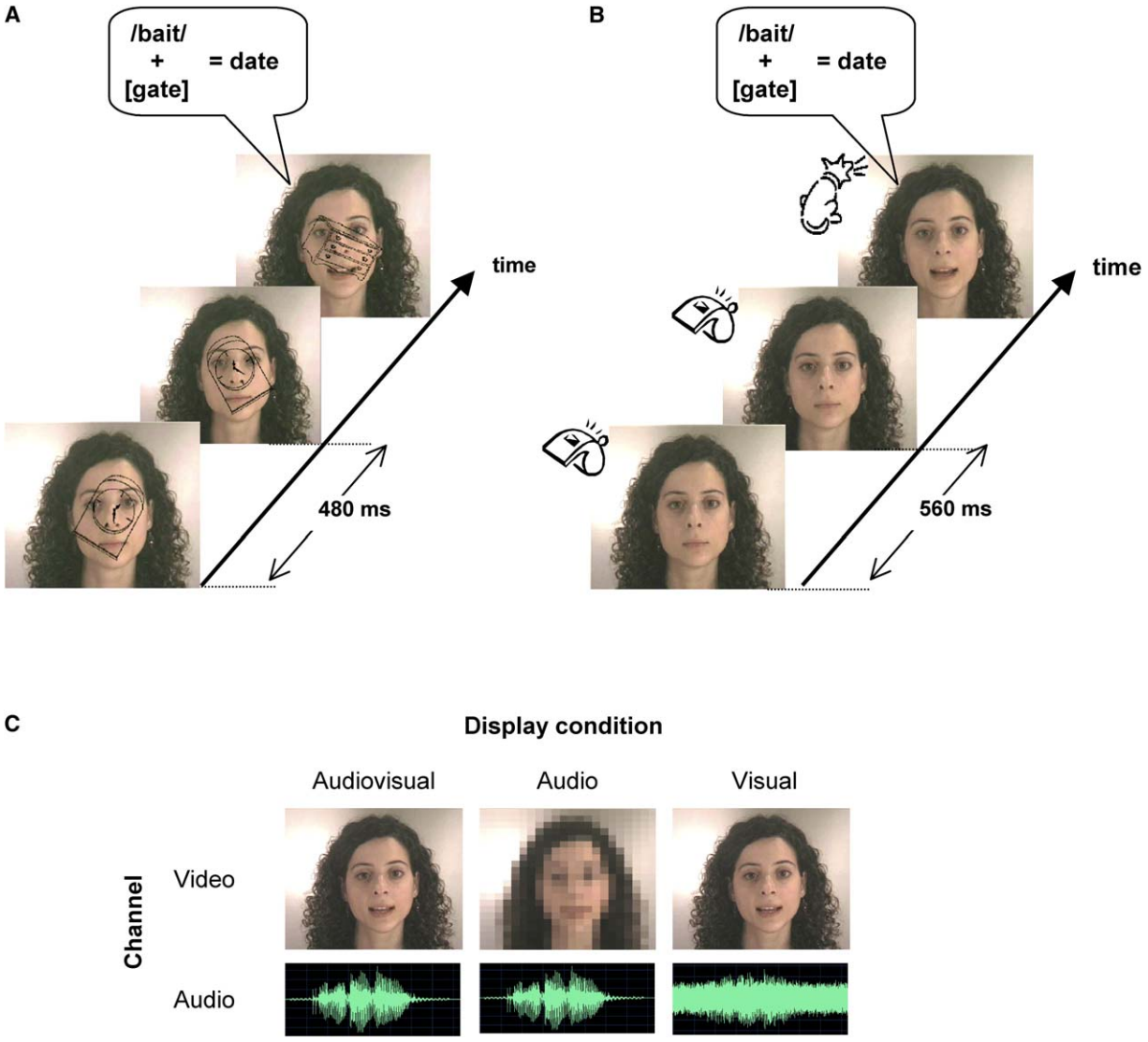

Figure 1. Experimental Conditions

The audiovisual word combinations leading to McGurk effect (see Supplemental Data) were presented at unpredictable moments (the interword interval was randomly distributed between 5 and 37 s). For each participant, every target in the word-recall task was presented only once throughout the experiment. A dual-task paradigm, consisting of a visual (Experiment 1) or auditory (Experiment 2) detection task, was used to divert attention from audiovisual blends (McGurk words). Half of the observers in each experiment detected repetitions in a concurrent stimulus stream in addition to recalling the audiovisual words (dual task), whereas the other half simply reported the words that the speaker said (single task). Targets in the repetition task occurred every six items on average and were not correlated with the words in the spoken stream.

(A) In Experiment 1, the repetition detection task was performed on a stream of line-drawn pictures superimposed on the video recording containing the speech material.

(B) In Experiment 2, the repetition detection task was performed on a rapid auditory stream (dual-task condition) of common sounds superimposed on the audio recording containing the speech material.

(C) Each participant performed the task (either single or dual) under three different conditions: audiovisual, auditory, and visual (the order was counterbalanced across participants and items). The visual-only condition was produced by adding white noise to the sound track (thus rendering the auditory words unintelligible). In the auditory-only conditions, a video quantization effect was applied to degrade the image to disrupt reliable vision of the lip movements while preserving the overall features of the video display (colors, motion, overall shape).

forming the concurrent repetition task did not interfere with word recall per se. That is, perception in each modality alone was not affected under divided attention, and yet the degree of integration between them was.

However, this result does not unequivocally imply that attention had an effect on audiovisual integration mechanisms. The observed reduction in visually influenced responses under dual-task conditions is equally compatible with the account that the demands of the visual distractor task prevented further processing of the visual speech information at an early processing stage, before audiovisual integration could take place.

Table 1. Proportion of Each Response Type (and SEM within Parentheses) as a Function of Task and Display Condition in Experiments 1 and 2

		Single Task			Dual Task		
	Condition	Audiovisual	Audio	Visual	Audiovisual	Audio	Visual
Response types in Experiment 1	Auditory	.61 (.07)	.93 (.02)	.01 (.00)	.87 (.02)	.91 (.02)	.00 (.00)
	Visual/fusion (McGurk)	.33 (.06)	.03 (.02)	.02 (.01)	.08 (.02)	.05 (.02)	.01 (.01)
	Other	.05 (.02)	.03 (.01)	.97 (.01)	.04 (.02)	.04 (.01)	.98 (.01)
Response types in Experiment 2	Auditory	.16 (.02)	.87 (.02)	.00 (.00)	.36 (.04)	.78 (.03)	.00 (.00)
	Visual/fusion (McGurk)	.81 (.03)	.09 (.02)	.07 (.02)	.58 (.04)	.12 (.01)	.06 (.02)
	Other	.02 (.01)	.04 (.01)	.93 (.01)	.06 (.02)	.10 (.03)	.93 (.02)

Note: For each experiment, the responses were classified according to whether they were visual/fusion responses (McGurk illusions; see main analyses in the text), auditory responses, or other. The table displays the average proportion (and SEM) of each type of response in each condition. In an additional ANOVA on the proportion of auditory responses, the findings reported in the main text were confirmed. In particular, in Experiment 1 there was a significant interaction between Task and Condition ($F_{Exp1} = 11.9$, $p = 0.002$), the interaction caused by the significant Task effect in the audiovisual condition ($t_{Exp1} = 3.5$, $p < 0.005$) but not in the other two conditions (both $|t| < 1$). In Experiment 2, there was a significant interaction between Task and Condition ($F_{Exp2} = 19.9$, $p < 0.001$), explained by the higher frequency of auditory responses under dual task than under single task in the audiovisual condition ($t_{Exp2} = 4.3$, $p = 0.001$), the null effect of task in the visual condition ($t = 1$), and a small but opposite effect in the auditory condition ($t_{Exp1} = 2.5$, $p < 0.05$). The ANOVA on the Other responses did not reveal any significant effects or interactions except that of Condition, whereby this kind of responses was more prevalent in the visual-alone condition than in the other two conditions ($p < 0.001$ in both experiments). Each average is based on 9 subjects × 26 trials, for a total of 234 observations per cell.

This interpretation is in agreement with previous reports suggesting that attention may have an effect on the processing of unimodal (visual) information rather than on the audiovisual binding process [13, 14]. If the interference observed under dual-task conditions reflects impairments at a modality-specific level, then a difficult concurrent task performed on an auditory stream should generate a reduction in auditory responses (i.e., increase in visually influenced responses). This would be in accord with the common observation that the McGurk illusion is more pronounced when the auditory signal is somewhat degraded [15]. However, if the hypothesis that audiovisual speech integration breaks down under high attention demands is true, then the prediction is different. The frequency of visually induced illusions in the audiovisual condition should *diminish* selectively under dual-task control in relation to single-task control.

Data from Experiment 2 (see Figure 3 for the group averages) showed that, again, the critical interaction between Task and Condition was significant ($F_1 = 12.9$, $p = 0.001$; $F_2 = 24.6$, $p < 0.001$). When each display condition was analyzed separately, only the audiovisual condition revealed an effect of Task, with a mean of 57.6% visual/fusion responses in the dual-task condition and 81.1% in single task ($t_1 = 4.5$, $p < 0.001$; $t_2 = 5.9$, $p < 0.001$). Thus, crucially, the magnitude of the McGurk effect decreased under dual-task conditions (by 23.5%), a pattern similar to that found in Experiment 1 (24.5% decrease). The control conditions (auditory-only and visual-only) did not reveal any significant effect of task on the proportion of visually influenced responses when tested separately ($t_1 = 1.2$, $p = 0.234$, $t_2 = 1.6$, $p = 0.118$ and $t_1 = 0.4$, $p = 0.71$, $t_2 = 0.4$, $p = 0.676$, respectively).

Here, we used a concurrent auditory processing task that, according to the prediction of the modality-specific interference hypothesis, should have degraded the perception of the auditory component of speech. Yet the results are clearly in the opposite direction. Increasing the demands on auditory attention in the dual-task group produced *more* auditory-based responses (i.e., less McGurk) in the audiovisual condition, in comparison to the single-task group (see Table 1). This finding is consistent with the hypothesis that exhausting attentional resources seriously compromises the multisensory integration process. It is also notable that, in comparison to Experiment 1, there was an overall increase of visual/fusion responses in the audiovisual condition, irrespective of task demands. This expected effect is

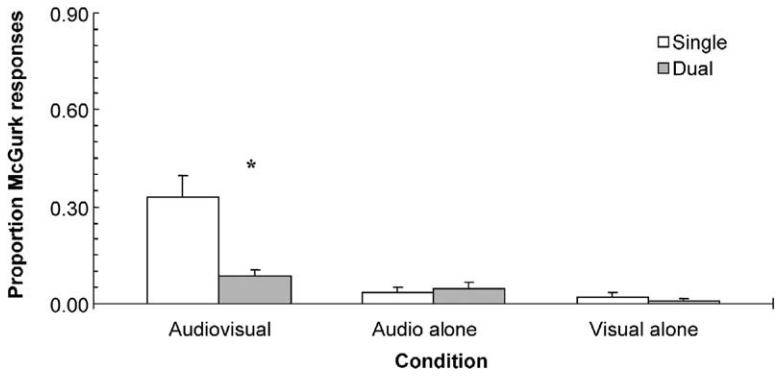

Figure 2. Results of Experiment 1

Experiment 1 showed that the percentage of fusion/visual responses in the audiovisual condition (McGurk effect) was significantly reduced when participants performed a concurrent visual task ($t = 3.8$, $p < .005$). The error bars represent the SEM. The asterisk denotes a significant difference between dual and single task.

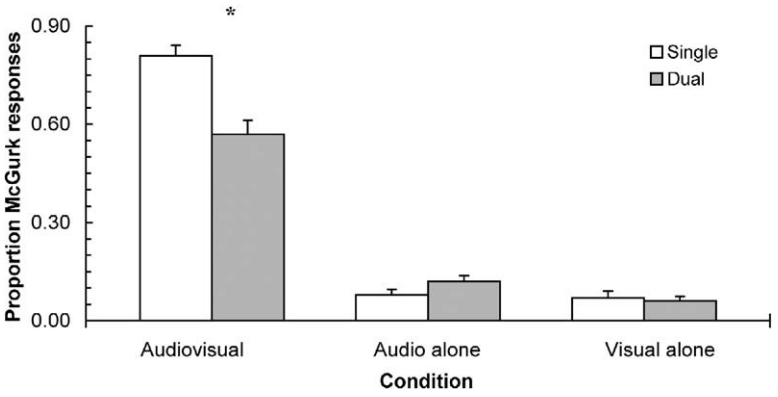

Figure 3. Results of Experiment 2

Experiment 2 showed that the percentage of fusion/visual responses in the audiovisual condition (McGurk effect) was significantly reduced when participants performed an auditory concurrent task ($t = 4.5$, $p < 0.001$). The error bars represent the SEM. The asterisk denotes a significant difference between dual and single task.

attributable to the partial masking from the concurrent-task sounds on the auditory words to be reported. That is, the overall increase in McGurk (i.e., visually dominated) responses can reflect a general enhancement of the visual influence on speech processing when intelligibility is compromised by noise [12]. Note also that the small increment in visual/fusion responses in the visual-only condition of Experiment 2 (as compared with Experiment 1) can be explained by the absence of line drawings on the speaker's face. However, adding to the strength of our conclusions, the selective reduction of audiovisual integration under dual-task conditions in Experiment 2 occurred over and above these standard effects of auditory noise and visual masking on audiovisual integration.

Our finding that audiovisual speech integration is susceptible to manipulations of attentional resources is especially relevant because it challenges previous attention-free accounts of the McGurk effect [3–8, 16], and it supports the claim that attention is necessary to bind features across modalities [17]. The strongest arguments supporting the preattentive nature of audiovisual speech integration rely on paradigms in which attention is directed to the speech event in the absence of other concurrent stimuli. In behavioral studies in which attention was manipulated either by explicitly instructing the observer to focus on a specific sensory modality [4, 5] or by diverting attention from the audiovisual stimuli indirectly [6], there is little effect on susceptibility to McGurk illusions. Similarly, human electrophysiological (ERP) measures of brain activity during McGurk stimulus perception do not suggest recruitment of attentional resources (in the form of a distinctive temporal or spatiotemporal electrophysiological signature) when audiovisual, rather than auditory, syllables are the target of processing [7, 8]. However, no experimental situation except for that described here has attempted to exhaust the amount of attentional resources available for audiovisual speech integration. According to attentional load theory [9], when the amount of resources required to perform a cognitive operation does not exceed the capacity of the system, the remaining attention resources may spill over to other processes even if irrelevant for the task. Because attentional demands in previous paradigms were relatively low, spare resources may have been devoted to the integration of audiovisual stimuli. In line with this idea, several recent studies now show that some perceptual phenomena classically considered as preattentive, such as visual-motion aftereffects, word reading, or parallel visual search, can indeed be modulated or even completely prevented when combined with a demanding concurrent task [10, 18–20]. Here, we have applied this logic for the first time to multisensory integration to reveal that audiovisual speech binding falters under high attention demands.

The finding that attention can modulate audiovisual integration is not necessarily in conflict with the neural evidence obtained with event-related potentials (ERP) or magnetoencefalography (MEG), which, to date, has explored audiovisual speech processing only in the context of low attention load. Under conditions of focused attention (no additional load), these previous studies suggest that audiovisual integration occurs at early processing stages [7, 21]. Critical to fMRI findings [15, 16] is the discovery that a multimodal processing region, the (posterior parts of) superior temporal sulcus (STS) within the superior temporal lobe, is preferentially activated by congruent audiovisual speech [22]. Not only is STS responsive to audiovisual speech "that fits" (including, we surmise, McGurk type stimulation), but also, activation here correlates with increased activity in sensory-specific regions (via "back projections") [23]. STS is itself extensively connected with both parietal and frontal systems implicated in discriminative functioning and the allocation of attention. This conceptualization is in line with recent studies investigating attentional influences on brain correlates of sensory processing outside the domain of speech. These have shown that not only can attention modulate higher-level (association) processing stages, but it can also have an influence at multiple levels of (early) sensory processing via feedback projections (producing shifts in baseline activity) [24, 25]. The fact that sensory processing can be strongly influenced by attentional mechanisms agrees well with the present results and allows one to reconcile neuroimaging evidence for early audiovisual binding and attentional modulation.

The literature on multisensory integration in humans contains several demonstrations that binding across modalities can occur in an automatic fashion [26, 27], independently of the focus of spatial attention in healthy

[28, 29] as well as brain-damaged observers [30], and that it can even serve as the basis to shift attention in space [31]. The present findings suggest, however, a limit to this automaticity. In the case of animal models, the effects of attention in multisensory integration have been less well documented. Yet animal electrophysiological studies reveal, for example, the critical role that projections from cortical association areas play in enabling the multisensory capabilities of certain polysensory sites even at the subcortical level (e.g., superior colliculus [32]; see [33] for a suggestion about a potential role of attention on multisensory integration in the SC).

Supplemental Data

Supplemental Results, detailed Experimental Procedures, and a supplemental table are available at http://www.current-biology.com/cgi/content/full/15/9/839/DC1/.

Acknowledgments

This research was supported by grants from the James McDonnell Foundation (JMCD20002079) and the Ministerio de Ciencia y Tecnología (Spain; TIN2004-04363-C03-02) and by a fellowship Beca de Formació en la Recerca i la Docència from the Universitat de Barcelona to A.A.

Received: November 23, 2004
Revised: March 8, 2005
Accepted: March 14, 2005
Published: May 10, 2005

References

1. Ghazanfar, A.A., and Logothetis, N.K. (2003). Facial expressions linked to monkey calls. Nature *423*, 937–938.
2. Burnham, D., and Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. Dev. Psychobiol. *45*, 204–220.
3. McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. Nature *265*, 746–748.
4. Massaro, D.W. (1987). Speech Perception by Ear and Eye (Hillsdale, NJ: LEA).
5. Dekle, D., Fowler, C., and Funnell, M. (1992). Auditory-visual integration in perception of real words. Percept. Psychophys. *51*, 355–362.
6. Soto-Faraco, S., Navarra, J., and Alsius, A. (2004). Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. Cognition *92*, B13–B23.
7. Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., and Deltenre, P. (2002). Mismatch negativity evoked by the McGurk–MacDonald effect: A phonetic representation within short-term memory. Clin. Neurophysiol. *113*, 495–506.
8. Bernstein, L.E., Auer, E.T., Jr., and Moore, J.K. (2004). Audiovisual speech binding: Convergence or association. In Handbook of Multisensory Processes, G.A. Calvert, C. Spence, and B.E. Stein, eds. (Cambridge, MA: MIT Press), pp. 203–224.
9. Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. J. Exp. Psychol. Hum. Percept. Perform. *21*, 451–468.
10. Joseph, J.S., Chun, M.M., and Nakayama, K. (1997). Attentional requirements in a "preattentive" feature search task. Nature *387*, 805–807.
11. Auer, E., and Bernstein, L.E. (1997). Speechreading and the structure of the lexicon: Modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. J. Acoust. Soc. Am. *102*, 3704–3710.
12. Sumby, W., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. *26*, 212–215.
13. Tiippana, K., Andersen, T.S., and Sams, M. (2004). Visual attention modulates audiovisual speech perception. Eur. J. of Cog. Psychol. *16*, 457–472.
14. Massaro, D.W. (1998). Perceiving Talking Faces: From Speech Perception to a Behavioral Principle (Cambridge, Massachusetts: MIT Press).
15. Sekiyama, K., Kanno, I., Miura, S., and Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. Neurosci. Res. *47*, 277–287.
16. Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P.K., Woodruff, P.W., Iversen, S.D., and David, A.S. (1997). Activation of auditory cortex during silent lipreading. Science *276*, 593–596.
17. Treisman, A., and Gelade, G. (1980). A feature-integration theory of attention. Cognit. Psychol. *12*, 97–136.
18. Rees, G., Frith, C.D., and Lavie, N. (1997). Modulating irrelevant motion perception by varying attentional load in an unrelated task. Science *278*, 1616–1619.
19. Rees, G., Russell, C., Frith, C.D., and Driver, J. (1999). Inattentional blindness versus inattentional amnesia for fixated but ignored words. Science *286*, 2504–2507.
20. Rees, G., Frith, C.D., and Lavie, N. (2001). Perception of irrelevant visual motion during performance of an auditory task. Neuropsychologia *39*, 937–949.
21. Möttönen, R., Schurmann, M., and Sams, M. (2004). Time course of multisensory interactions during audiovisual speech perception in humans: A magnetoencephalographic study. Neurosci. Lett. *363*, 112–115.
22. Calvert, G.A., Campbell, R., and Brammer, M.J. (2000). Evidence from functional magnetic resonance imaging of cross-modal binding in the human heteromodal cortex. Curr. Biol. *10*, 649–657.
23. Calvert, G.A., Brammer, M.J., Bullmore, E.T., Campbell, R., Iversen, S.D., and David, A.S. (1999). Response amplification in sensory-specific cortices during cross-modal binding. Neuroreport *10*, 2619–2623.
24. Driver, J., and Frith, C. (2000). Shifting baselines in attention research. Nat. Rev. Neurosci. *1*, 147–148.
25. Driver, J., and Spence, C. (2000). Multisensory perception: Beyond modularity and convergence. Curr. Biol. *10*, R731–R735.
26. Bertelson, P., and Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. Psychon. Bull. Rev. *5*, 482–489.
27. Caclin, A., Soto-Faraco, S., Kingstone, A., and Spence, C. (2002). Tactile 'capture' of audition. Percept. Psychophys. *64*, 616–630.
28. Vroomen, J., Driver, J., and de Gelder, B. (2001). Is cross-modal integration of emotional expressions independent of attentional resources? Cogn. Affect. Behav. Neurosci. *1*, 382–387.
29. 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.
30. Bertelson, P., Pavani, F., Ladavas, E., Vroomen, J., and de Gelder, B. (2000). Ventriloquism in patients with unilateral visual neglect. Neuropsychologia *38*, 1634–1642.
31. Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. Nature *381*, 66–68.
32. Stein, B., Stanford, T., Wallace, M., Vaughan, J., and Jiang, W. (2004). Crossmodal spatial interactions in subcortical and cortical circuits. In Crossmodal Space and Crossmodal Attention, C. Spence and J. Driver, eds. (Oxford: Oxford University Press), pp. 243–264.
33. Populin, L.C., and Yin, T.C.T. (2002). Bimodal interactions in the superior colliculus of the behaving cat. J. Neurosci. *22*, 2826–2834.