Please cite this article in press as: Munhall et al., Audiovisual Integration of Speech in a Bistable Illusion, Current Biology (2009), doi:10.1016/j.cub.2009.03.019

Current Biology *19*, 1–5, May 12, 2009 ©2009 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2009.03.019

Report

Audiovisual Integration of Speech in a Bistable Illusion

K.G. Munhall,[1,2,*] M.W. ten Hove,[3] M. Brammer,[5] and M. Paré[1,4]
[1]Department of Psychology
[2]Department of Otolaryngology
[3]Department of Ophthalmology
[4]Department of Physiology
Queen's University
Kingston, ON K7L 3N6
Canada
[5]Centre for Neuroimaging Sciences
Institute of Psychiatry
London SE5 8AF
UK

Summary

Visible speech enhances the intelligibility of auditory speech when listening conditions are poor [1], and can modify the perception of otherwise perfectly audible utterances [2]. This audiovisual perception is our most natural form of communication and one of our most common multisensory phenomena. However, where and in what form the visual and auditory representations interact is still not completely understood. Although there are longstanding proposals that multisensory integration occurs relatively late in the speech-processing sequence [3], considerable neurophysiological evidence suggests that audiovisual interactions can occur in the brain stem and primary sensory cortices [4, 5]. A difficulty testing such hypotheses is that when the degree of integration is manipulated experimentally, the visual and/or auditory stimulus conditions are drastically modified [6, 7]; thus, the perceptual processing within a modality and the corresponding processing loads are affected [8]. Here, we used a bistable speech stimulus to examine the conditions under which there is a visual influence on auditory perception in speech. The results indicate that visual influences on auditory speech processing, at least for the McGurk illusion, necessitate the conscious perception of the visual speech gestures, thus supporting the hypothesis that multisensory speech integration is not completed in early processing stages.

Results and Discussion

In the present studies, we held audiovisual stimulus conditions constant and allowed subjective organization of the percept to determine the extent of multisensory integration. This was achieved through the use of a dynamic version of Rubin's vase illusion [9]. In our stimulus, an irregular vase rotated and its changing profile produced a talking face profile (see Figure 1). The face articulated the nonsense utterance /aba/ while the accompanying acoustic signal was a voice saying the nonsense utterance /aga/. Two visual and two auditory percepts occur with these stimuli. Visually, the faces appeared to be the figure and the vase was the background or vice versa. Auditorily, subjects heard either the recorded audio track, /aga/, or heard the so-called combination McGurk effect, /abga/ [3]. In this illusion, both consonants are "heard" even though only the /g/ is present in the acoustic signal. This percept results from visual influences on auditory perception. When subjects only heard the acoustic signal, /aga/, there was no phonetic influence of the visual information. Three experiments are presented here.

Experiment 1 looked at the association of the McGurk illusion and the perception of either the vase or the face. Complete independence of these percepts would suggest that visual influences on auditory speech perception might occur at an early stage of processing, either subcortically or in the primary sensory cortex. Recent work on figure-ground perception indicates that, beyond the simple competition between low-level processing units, figural assignment may involve widespread recurrent processing (e.g., [10]) and biased competition between high-level shape perception units [11]. If audiovisual integration in speech is not sensitive to the suppression of face perception in the bistable stimulus, it must precede or be independent of this process. Alternatively, complete association of face perception and perception of the McGurk illusion would suggest that audiovisual integration of speech depended on categorical stimulus representations for object perception. Two different stimuli were presented to subjects. In the first condition, the vase rotated, and its shape produced a profile of an articulating face saying the utterance /aba/ (Figure 1A: moving face, moving vase). In the second condition, the vase rotated, but the face profile remained constant (Figure 1B: still face, moving vase). This was achieved by subtle changes to the three-dimensional (3D) vase in this condition such that its visible rotation did not produce any profile changes. Such a stimulus could only be produced by using animation. Each of these stimuli was combined with a recording of /aga/. The control condition was not expected to produce the McGurk effect since there was no visual information for a consonant. Subjects watched single tokens and gave two responses. First, they reported whether they perceived a vase or a face, then they told the experimenter whether they heard /aga/ or /abga/.

For the moving face, moving vase stimulus, the results show a strong association between consciously perceiving the face and perceiving the McGurk effect (Figure 2); 66% of the responses shared this perceptual pattern. Only 9% of the responses reported the McGurk effect when the vase was the percept. The control stimulus (still face, moving vase) produced a quite different pattern of responses. Approximately 90% of the speech responses were percepts of the auditory stimulus /aga/. These responses were split between the vase and face percepts, with a slight bias toward perceiving the face. The /abga/ responses (~10%) were split between the face and vase percepts. This three-way interaction was reliable as assessed by the chi-square test ($p < 0.001$). When the 2×2 response contingencies tables were evaluated separately for each stimulus, the moving face, moving vase stimulus showed a reliable association between face perception and the perception of the McGurk

*Correspondence: kevin.munhall@queensu.ca

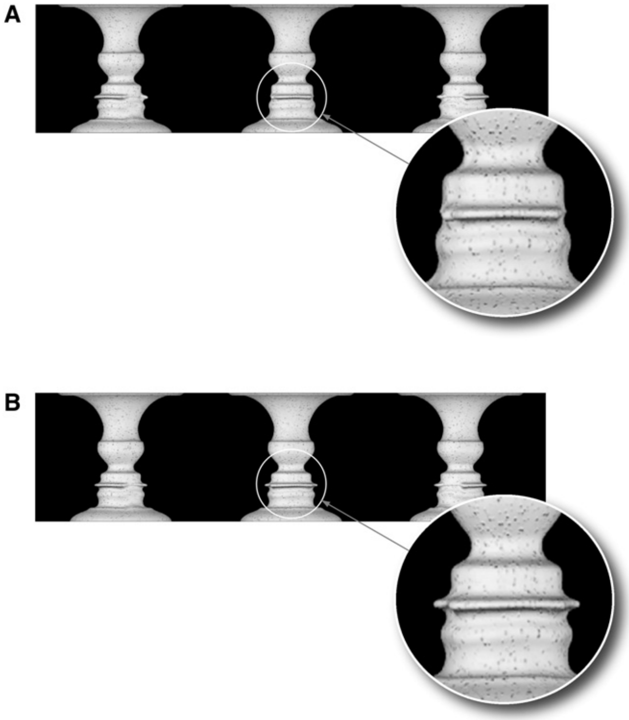

Figure 1. Individual Frames from the Rotating Vase Movie Used in the Dynamic Vase Conditions in Experiments 2 and 3

(A) Moving face, moving vase: the face profile changes shape as if the face is articulating the utterance /aba/ as the vase rotates. The three frames correspond to the points in time during the first vowel, during the closure for the /b/, and during the second vowel. The circle shows the detail of the lip closure in the facial profile.

(B) Still face, moving vase: the face profile does not change shape as the vase rotates. The three frames correspond to the same points in time as the sequence shown in (A). The circle shows the detail of the open lips in this condition, in contrast to that shown in (A).

combination (Fisher's exact probability test, $p < 0.05$), whereas the stimulus with a still face and moving vase showed no association ($p > 0.5$).

The small number of /bg/ percepts in the moving face, moving vase condition when the vase was reported was approximately equal to the number of /bg/ percepts for the still face, moving vase condition (∼10%). This common response rate suggests that this may be simply response bias or error. Although motion in a suppressed image in binocular rivalry can still elicit motion aftereffects [12] and contribute to the perception of apparent motion [13], the moving face seems to require conscious perception in order to influence auditory speech.

The presence of vase motion alone produced a large number of face percepts. This is not associated with audiovisual integration, given that virtually no McGurk effects were observed for this condition. When the two motion conditions in experiment 1 are contrasted, we see strong evidence for the importance of dynamic facial information and its conscious perception as prerequisites for audiovisual speech perception. These findings are consistent with those from studies showing that awareness that an acoustic signal is speech is a prerequisite for audiovisual integration [14].

Two control experiments were carried out to help clarify the results. Experiment 2 tested further how motion influenced vase/face perception and, in addition, how sound influenced this percept. Three different levels of movement of the stimulus were shown with and without the speech soundtrack. In one condition, a static frame of the vase and the face was shown for the duration of the dynamic stimuli. This frame was identical to the leftmost frame in Figure 1A. The other two conditions were identical to the visual conditions tested in experiment 1.

Figure 3 shows the mean proportion of face percepts for the three movement conditions as a function of whether a speech utterance was played along with the visual stimuli. A robust effect of movement condition is evident [$F(2,24) = 36.4$, $p < 0.001$], whereas only a modest influence of the presence of sound ($F(1, 24) = 3.8$, $p = 0.06$), and no interaction between sound and movement conditions can be seen. The presence of motion dramatically decreased the percentage of vase percepts from the high of 76% in the static image condition to a low of 28% in the moving face, moving vase stimulus. Each of the three motion conditions was reliably different from one another ($p < 0.01$). The presence of auditory speech increased the percentage of face percepts, but by less than 10% on average.

From a pictorial viewpoint, the stimulus was biased toward perceiving the vase by the surface texture information and 3D rendering of the vase [15]. The still image's high proportion of vase percepts reflects this. When auditory speech and any motion (either the vase alone or both the vase and the face) were presented, the proportion of vase percepts consistently decreased from the silent, still image-condition high-water mark. The onset of motion in an image during binocular rivalry [16] or higher velocity [17] in an image tends to increase the likelihood of that image dominating perception. The reduction in vase percepts in the still face, moving vase condition is inconsistent with these findings. The independence of facial form and motion pathways [18] suggests a possible high-level associative account. Nevertheless, the moving face, moving vase condition is the only visual condition in which face percepts dominate (>50%). The influence of sound was modest and relatively consistent across the different visual conditions. If early audiovisual interactions were driving the visual percepts, the moving face condition would have been expected to be the condition most strongly influenced by the presence of sound. This was not the case. The results suggest that the conditions determining the perception of the unimodal stimulus (vision) are primarily determining multisensory integration [19].

Experiment 3 was carried out to test whether perceptual alternations could be accounted for simply by an alternation between eccentric (face) and central (vase) fixations. The distributions of gaze fixation positions associated with either of the two reported percepts were compared with an analysis derived from signal-detection theory. For each subject, we found that the distribution of fixation positions associated with each percept overlapped extensively, and only in 0.1% of the cases (22/23506) could the gaze distributions be considered as significantly different. This finding is consistent with the report that the changes in perception of the Rubin's face-vase stimulus are not associated with changes in eye positions [20] and with work showing that the McGurk effect is not dependent on whether the visual speech is viewed by using central (foveal) or paracentral vision (e.g., [21]).

Recent evidence indicates that the attentional state of the subject influences audiovisual integration of speech [8, 22]. The McGurk effect is reduced under high-attention demands.

Please cite this article in press as: Munhall et al., Audiovisual Integration of Speech in a Bistable Illusion, Current Biology (2009), doi:10.1016/j.cub.2009.03.019

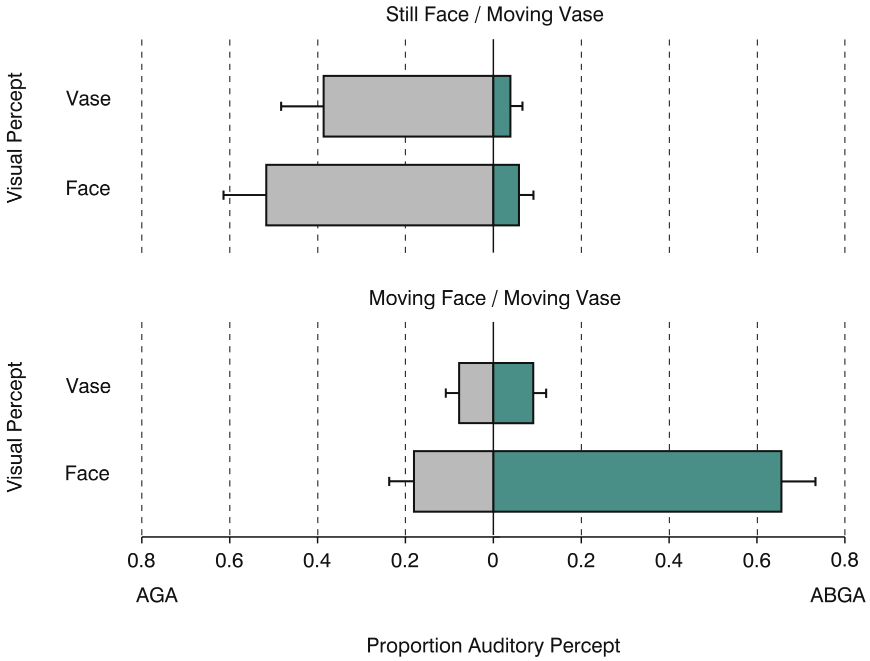

Figure 2. Proportion of Different Speech Percepts as a Function of Whether the Face or Vase Was Perceived for the Still Face, Moving Vase and the Moving Face, Moving Vase Conditions

The proportions are computed separately for the two stimulus conditions (still face, moving vase [upper panel]; moving face, moving vase [lower panel]) and sum to 1.0 for each panel. AGA responses correspond to perception of the sound track, whereas ABGA responses indicate the McGurk combination percept. The error bars correspond to the standard errors of the means.

Furthermore, subjects appear to have perceptual access to the individual sensory components as well as the unified multisensory percept [23]. Findings such as these contradict the view that multisensory integration is preattentive and thus automatic and mandatory [24] and are consistent with the involvement of higher-order processes in phonetic decisions. The evidence that auditory processing is influenced by visual information subcortically as early as 11 ms after acoustic onset for speech stimuli [4] or cortically in less than 50 ms [5] for tone stimuli is, at first look, difficult to reconcile with such findings.

One possible solution is that multisensory speech processing involves an interaction between auditory and visual information at many levels of perception, yet the final phonetic categorization, and ultimately audiovisual integration, takes place quite late. Multisensory processing may involve rapid attentional mechanisms that modulate early auditory or visual activity [25], promote spatial orienting [26], or provide contextual modulation of activity [27]. Yet, the dynamic structure of speech may require integration over longer timescales than the speed at which vision and audition can initially interact. The production of human speech is quite slow, with the modal syllable rate being ~3–6 Hz [28]. It has long been recognized that information for speech sounds does not reside at any instant in time, but rather is extended over the syllable [29]. Thus, even within a modality the temporal context of information determines its phonetic identity. For audiovisual speech of the kind presented here, the information for consonant identity is extended in time [30], and perception requires extended processing to integrate this perceptual information.

It remains to be seen whether this conclusion extends to all audiovisual speech phenomena. Vision can influence auditory speech perception in at least two distinct ways [31]. The first involves correlational modulation. Visible speech strongly correlates with some parts of the acoustic speech signal [32]. The acoustic amplitude envelope and even the detailed acoustic spectrum can be predicted by the visible speech articulation. This redundancy may permit early modulation of audition by vision, for example, by the visual signal amplifying correlated auditory inputs [33].

The second way in which visible speech influences auditory speech is by providing complementary information. In this case, vision provides stronger cues than the auditory signal or even information missing from the auditory signal. This latter case is the situation that best describes the perception of speech in noise and the combination McGurk effect. In both of the examples, the correlation between auditory and visual channels is broken because of the loss of information in the auditory channel. For the combination McGurk,

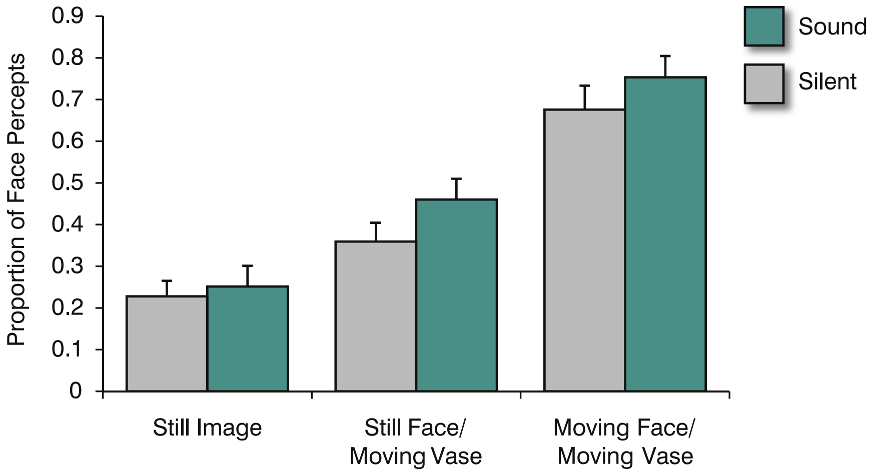

Figure 3. Proportion of Vase Percepts Reported as a Function of Visual Motion Conditions and the Presence or Absence of Auditory Speech

The error bars correspond to the standard errors of the means.

a /b/ could be plausibly produced during the intervocalic closure in /aga/ with minimal or without any auditory cues. The strong cue of a visible bilabial closure provides independent information to the speech system. It is possible that such complementary visual information can only be combined with the auditory signal late in the phonetic decision-making process after both modalities carry out considerable processing.

In experimental settings, the natural correlation between auditory and visual speech can also be broken by having the visible speech provide contradictory cues for the auditory signal. This is the case for the standard fusion McGurk effect [2], in which an auditory /b/ is combined with a visual /g/ and /d/ is heard. Both modalities yield sufficient but contradictory cues for consonant perception, although for the strongest effect the auditory /b/ must be a weak percept. Whether the perceptual system also makes a late phonetic decision under these conditions is unclear. The evidence from attention studies suggests that this is the case [8, 22].

Bistable phenomena in vision, audition, and multisensory processing are well accounted for by ideas of distributed competition involving different neural substrates and perceptual processes [34–36]. Audiovisual speech perception may share this form of distributed processing. However, the data presented here indicate that multisensory decision making in speech perception requires high-level phonetic processes including the conscious perception of facial movements. The unique stimuli used in these experiments will be an important tool in further characterizing the network of processes involved in this multisensory perception.

Experimental Procedures

Subjects

The studies were approved by the Queen's University General Research Ethics Board, and all subjects gave informed consent before participating in the research.

Stimuli

We created audiovisual stimuli by using a dynamic version of the Rubin Vase illusion [9]. Experiments 1 and 2 used an animated version (Figure 1) of a vase created with Maya (Autodesk), and the face profile was determined by the same video sequence used to create the stimuli used in experiment 3. In both stimuli, the vase was irregular, and as it rotated, its edge would produce a different face profile. Figure 1A shows three frames from the movie in which the vase rotates, and its changing shape produced a face profile that articulates the utterance /aba/. The face profile matches the original movie exactly on a frame-by-frame basis. Figure 1B shows three frames from the control movie in which a slightly different vase rotates; however, its changing shape produces no change in the face profile. The difference in profile changes between Figure 1A and Figure 1B is due to subtle differences in the animated 3D vase shape between the two conditions. In experiment 3, a video of a rotating, custom-constructed vase was edited with the profile of a female speaker saying the utterance /aba/.

Procedure

Experiment 1

A total of 12 subjects were presented with two types of stimuli in single trials (a rotating vase that produced an articulating face, a rotating vase that produced a still face). Both were presented with the audio track /aga/. Subjects were asked to indicate whether they saw a face or a vase. Only a single response was permitted for each trial. After reporting what they saw, subjects were instructed to record whether the sound they perceived was most like /aga/ or /abga/. After ten warmup trials, the subjects were presented with 60 experimental trials, 30 of each condition in randomized order.

Experiment 2

A total of 14 subjects were presented with six types of stimuli in single trials. Three visual stimuli (still frame, moving face, moving vase: rotating vase that produced an articulating face, still face, moving vase: rotating vase that produced a still face) were presented with either the audio track /aga/ or silence. Each trial was composed of a single rotation of the vase or, in the case of the still frame, a period equaling the duration of the dynamic trials. The subjects' task was to indicate whether they saw the face or the vase first, then indicate each time it changed within a trial. After 12 warmup trials, each of the stimuli was presented five times with order randomized across stimulus type, comprising 30 experimental trials in total. When subjects reported more than one state in a single trial both responses were included in the analyses as separate responses for that condition. Subjects generally reported only one perceptual state for the bistable stimulus in each trial, and the overall average number of states reported was 1.05 states per trial. There were no differences in the number of states seen across the conditions.

Experiment 3

We tested seven subjects on a behavioral task in which the stimulus was displayed in loops of ten continuous utterances. Subjects responded after each loop with a key press indicating whether they heard a /b/ sound or not. In addition to the behavioral task, we examined whether the varying audiovisual percept of the bistable stimulus depended on the subject's gaze fixation position by monitoring the horizontal and vertical eye position of the subjects while they viewed the stimulus during repeated trials and reported when their percept changed.

We sampled horizontal and vertical eye positions at a rate of 1 kHz by using the search-coil-in-magnetic-field technique [37] with an induction coil that consisted of a light coil of wire embedded in a flexible ring of silicone rubber (Skalar) that adheres to the limbus of the human eye, concentric with the cornea [38]. The search coil was positioned in the dominant eye of the subjects after the surface of the eye had been anesthetized with a few drops of anesthetic (Tetracaine HCl, 0.5%). Details of this method were described previously [21].

Analysis: Experiment 3

The distributions of fixations for the signal-detection analysis were computed in the following manner: At each millisecond in each type of utterance (i.e., ones perceived either as /bg/ or /g/), we calculated separately the probabilities that the positions of horizontal and vertical gaze fixation were greater than a position criterion, which was incremented in 1° steps across the image from either the left margin of the image or its bottom margin. The ensuing fixation position probabilities (for each percept) were then plotted against each other in a receiver operating characteristic (ROC) curve, and the area under each curve (AUROC) was computed to capture the amount of separation between the two distributions of fixation positions. This quantitative measure gives the general probability that, given one draw from each distribution, the fixation positions from the distributions associated with the two percepts are distinct.

Acknowledgments

The research was funded by the National Institute on Deafness and Other Communication Disorders (grant DC-00594) and by the Natural Sciences and Engineering Research Council of Canada. Bryan Burt assisted in data collection. The authors acknowledge the artistic work of Patty Petkovich, Rob Loree, and particularly Mike Watters for creation of the vase stimuli.

Received: November 21, 2008
Revised: March 1, 2009
Accepted: March 2, 2009
Published online: April 2, 2009

References

1. Sumby, W.H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. *26*, 212–215.
2. McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing speech. Nature *264*, 746–748.
3. Massaro, D.W. (1998). Perceiving Talking Faces: From Speech Perception to a Behavioral Principle (Cambridge, MA: MIT Press).
4. Musacchia, G., Sams, M., Nicol, T., and Kraus, N. (2006). Seeing speech affects information processing in the human brainstem. Exp. Brain Res. *168*, 1–10.
5. Molholm, S., Ritter, W., Murray, M.M., Javitt, D.C., Schroeder, C.E., and Foxe, J.J. (2002). Multisensory auditory–visual interactions during early sensory processing in humans: a high-density electrical mapping study. Brain Res. Cogn. Brain Res. *14*, 115–128.

6. Ross, L.A., Saint-Amour, D., Leavitt, V.N., Javitt, D.C., and Foxe, J.J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cereb. Cortex *17*, 1147–1153.
7. Callan, D.E., Jones, J.A., Munhall, K.G., Kroos, C., Callan, A., and Vatikiotis-Bateson, E. (2004). Multisensory-integration sites identified by perception of spatial wavelet filtered visual speech gesture information. J. Cogn. Neurosci. *16*, 805–816.
8. Alsius, A., Navarra, J., Campbell, R., and Soto-Faraco, S.S. (2005). Audiovisual integration of speech falters under high attention demands. Curr. Biol. *15*, 839–843.
9. Rubin, E. (1915). Synoplevde Figurer (Kopenhagen: Gyldendalske).
10. Domijan, D., and Setic, M. (2008). A feedback model of figure-ground assignment. J. Vis. *8*, 1–27.
11. 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.
12. Lehmkuhle, S., and Fox, R. (1976). Effect of binocular rivalry suppression on the motion aftereffect. Vision Res. *15*, 855–859.
13. Wiesenfelder, H., and Blake, R. (1991). Apparent motion can survive binocular rivalry suppression. Vision Res. *31*, 1589–1600.
14. Tuomainen, J., Andersen, T., Tiippana, K., and Sams, M. (2005). Audiovisual speech perception is special. Cognition *96*, B13–B22.
15. Hasson, U., Hendler, T., Bashat, D.B., and Malach, R. (2001). Vase or face? A neural correlate of shape-selective grouping processes in the human brain. J. Cogn. Neurosci. *13*, 744–753.
16. Fox, R., and Check, R. (1968). Detection of motion during binocular rivalry suppression. J. Exp. Psychol. *78*, 388–395.
17. Blake, R., Yu, K., Lokey, M., and Norman, H. (1998). Binocular rivalry and visual motion. J. Cogn. Neurosci. *10*, 46–60.
18. Alais, D., and Parker, A. (2006). Independent binocular rivalry processes for form and motion. Neuron *52*, 911–920.
19. Sanabria, D., Soto-Faraco, S., Chan, J., and Spence, C. (2005). Intramodal perceptual grouping modulates multisensory integration: evidence from the crossmodal dynamic capture task. Neurosci. Lett. *377*, 59–64.
20. Andrews, T.J., Schluppeck, D., Homfray, D., Matthews, P., and Blakemore, C. (2002). Activity in the fusiform gyrus predicts conscious perception of Rubin's vase-face illusion. Neuroimage *17*, 890–901.
21. Paré, M., Richler, R., ten Hove, M., and Munhall, K.G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. Percept. Psychophys. *65*, 553–567.
22. Tiippana, K., Andersen, T.S., and Sams, M. (2004). Visual attention modulates audiovisual speech perception. Eur. J. Cogn. Psychol. *16*, 457–472.
23. Soto-Faraco, S., and Alsius, A. (2007). Conscious access to the unisensory components of a cross-modal illusion. Neuroreport *18*, 347–350.
24. 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.
25. Foxe, J.J., Simpson, G.V., and Ahlfors, S.P. (1998). Cued shifts of intermodal attention: parieto-occipital |10 Hz activity reflects anticipatory state of visual attention mechanisms. Neuroreport *9*, 3929–3933.
26. Spence, C., and Driver, J. (1997). Audiovisual links in exogenous covert spatial orienting. Percept. Psychophys. *59*, 1–22.
27. Kayser, C., Petkov, C.I., and Logothetis, N.K. (2008). Visual modulation of neurons in auditory cortex. Cereb. Cortex *18*, 1560–1574.
28. Greenberg, S. (1999). Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation. Speech Commun. *29*, 159–176.
29. Liberman, A.M., Cooper, F.S., Shankweiler, D.P., and Studdert-Kennedy, M. (1967). Perception of the speech code. Psychol. Rev. *74*, 431–461.
30. Munhall, K.G., and Tohkura, Y. (1998). Audiovisual gating and the time course of speech perception. J. Acoust. Soc. Am. *104*, 530–539.
31. Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. Philos. Trans. R. Soc. Lond. B Biol. Sci. *363*, 1001–1010.
32. Yehia, H.C., Kuratate, T., and Vatikiotis-Bateson, E. (2002). Linking facial animation, headmotion and speech acoustics. J. Phonetics *30*, 555–568.
33. Schroeder, C.E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (2008). Neuronal oscillations and visual amplifications of speech. Trends Cogn. Sci. *12*, 106–113.
34. Pressnitzer, D., and Hupé, J.-M. (2006). Temporal dynamics of auditory and visual bistability reveal principles of perceptual organization. Curr. Biol. *16*, 1351–1357.
35. Blake, R., and Logothetis, N.K. (2002). Visual competition. Nat. Rev. Neurosci. *3*, 13–21.
36. Shams, L., Kamitani, Y., and Shimojo, S. (2002). Visual illusion induced by sound. Brain Res. Cogn. Brain Res. *14*, 147–152.
37. Robinson, D.A. (1963). A method of measuring eye movements using a scleral search coil in a magnetic field. IEEE Trans. Biomed. Eng. *10*, 137–145.
38. Collewijn, H., van der Mark, F., and Jansen, T.C. (1975). Precise recording of human eye movements. Vision Res. *15*, 447–450.