When half a face is as good as a whole: Effects of simple substantial occlusion on visual and audiovisual speech perception

Timothy R. Jordan · Sharon M. Thomas

Abstract The visible movement of a talker's face is an influential component of speech perception. However, the ability of this influence to function when large areas of the face (~50%) are covered by simple substantial occlusions, and so are not visible to the observer, has yet to be fully determined. In Experiment 1, both visual speech identification and the influence of visual speech on identifying congruent and incongruent auditory speech were investigated using displays of a whole (unoccluded) talking face and of the same face occluded vertically so that the entire left or right hemiface was covered. Both the identification of visual speech and its influence on auditory speech perception were identical across all three face displays. Experiment 2 replicated and extended these results, showing that visual and audiovisual speech perception also functioned well with other simple substantial occlusions (horizontal and diagonal). Indeed, displays in which entire upper facial areas were occluded produced performance levels equal to those obtained with unoccluded displays. Occluding entire lower facial areas elicited some impairments in performance, but visual speech perception and visual speech influences on auditory speech perception were still apparent. Finally, implications of these findings for understanding the processes supporting visual and audiovisual speech perception are discussed.

Keywords Speech perception · Face perception · Hearing · Audiovisual speech · McGurk effect

T. R. Jordan (✉) · S. M. Thomas
College of Medicine, Biological Sciences and Psychology,
University of Leicester,
Henry Wellcome Building,
Leicester LE1 9HN, UK
e-mail: Prof.TimJordan@gmail.com

Introduction

The facial movements that accompany speech production (*visual speech*) are a powerful component of speech perception (e.g., Erber, 1969; Jordan, McCotter, & Thomas, 2000; MacLeod & Summerfield, 1987, 1990; McGurk & MacDonald, 1976; Middleweerd & Plomp, 1987; Reisberg, McLean, & Goldfield, 1987; Rosenblum, Yakel, & Green, 2000; Sumby & Pollack, 1954; Thomas & Jordan, 2002). In particular, seeing the articulating face of a talker can improve auditory speech intelligibility substantially in quiet and noisy environments, and in the McGurk effect (after McGurk & MacDonald, 1976), can alter the perceived identity of speech sounds. However, although these effects are well-established, the facial information required for perception of visual speech has yet to be fully revealed.

Several studies have helped to determine this facial information by leaving visible only an individual facial feature (see, e.g., Benoit, Guiard-Marigny, Le Goff, & Adjoudani, 1996; Berger, Garner, & Sudman, 1971; Cohen, Walker, & Massaro, 1996; Greenberg & Bode, 1968; IJsseldijk, 1992; Larr, 1959; Marassa & Lansing, 1995; McGrath, 1985; Montgomery & Jackson, 1983; Stone, 1957; Summerfield, 1979; Summerfield, MacLeod, McGrath, & Brooke, 1989; Summerfield & McGrath, 1984). For example, Summerfield (1979) presented displays in which the talker's lips were coated with ultraviolet paint so that only the lips could be seen. These lips-only displays produced a significant increase in auditory speech recognition compared to when no visual information was presented, although the increase was greater when whole faces were shown (see also Rosenblum, Johnson, & Saldaña, 1996). IJsseldijk (1992) found that visual speech perception was slightly poorer for displays in which only the mouth was visible compared to

whole-face displays (see also Berger et al., 1971; Greenberg & Bode, 1968; Marassa & Lansing, 1995), and Stone (1957) found no difference in visual speech perception between whole-face displays and displays in which only the mouth could be seen.

However, the tolerance of visual speech perception towards loss of facial information is far from understood. In particular, while previous studies have focused on maintaining information from an individual facial feature (e.g., the lips or mouth), a natural system of visual and audiovisual speech perception is likely to develop to cope with everyday occlusions that do not obscure all of a face except for the precise parameters of a particular feature. Indeed, faces in everyday environments are naturally obscured simply and extensively in various uncontrolled ways, by intervening objects, other people, shadows, the talker's own hand or hair, and so on. These situations leave the observer with views of a talker's face in which a substantial area of the face is fully occluded by an intervening surface without leaving visible only an area of the face that conveniently and specifically corresponds to the precise location and extent of a particular facial feature (we call these *simple substantial occlusions*). In contrast, occluding all of a talker's face except for the precise parameters of a particular feature (e.g., only the lips or mouth) would require a highly specific (and implausible) set of environmental circumstances that are unlikely to fully demonstrate the influence of occlusion on visual speech perception. Indeed, revealing only a feature in a talking face is likely to induce an abnormal focus of visual and attentional resources that may exaggerate the feature's influence on visual speech perception and distort an understanding of the influence of other areas of the face. Thus, while studying effects of occlusion on visual speech perception by carefully revealing only isolated features of the face is important, visual speech stimuli covered by simple substantial occlusions will help provide a more complete indication of the natural tolerance of visual speech perception towards the loss of information produced by facial occlusion.

The possibility that visual and audiovisual speech perception is tolerant of simple substantial occlusions is strengthened by the highly correlated movement of different areas of the face during articulation. For example, Munhall and Vatikiotis-Bateson (1998; see also Vatikiotis-Bateson, Munhall, Hirayama, Lee, & Terzopoulos, 1996; Yehia, Kuratate, & Vatikiotis-Bateson, 2002; Yehia, Rubin, & Vatikiotis-Bateson, 1998) showed that the shape and movement of the oral area produced during a range of articulations correlated at .95 or better with the movement produced by the same articulations in the outer regions of the face (the side of the jaw and cheeks). This suggests that movements in a range of different facial areas can provide useful

information for visual and audiovisual speech perception even when areas of prominent movement (especially the mouth) cannot be seen (e.g., Davis & Kim, 2006; Greenberg & Bode, 1968; Larr, 1959; Preminger, Lin, Payen, & Levitt, 1998; Scheinberg, 1980; Stone, 1957; Thomas & Jordan, 2004). Indeed, Scheinberg (1980) reported that observers could use "cheek-puffiness" to distinguish between consonants that look very similar at the mouth, and Preminger et al. (1998) also reported that the cheeks were important extraoral facial areas for visual speech perception. Moreover, visual speech perception occurs even when the mouth is static, or is digitally replaced by adjacent skin, in a fully visible and otherwise normal talking face (Thomas & Jordan, 2004) and may even be extracted from the top of the head (Davis & Kim, 2006). The indications are, therefore, that visual speech can be encoded from a range of locations and, consequently, that visual and audiovisual speech perception may survive a range of simple substantial occlusions remarkably well.

However, factors other than the distribution of visual speech across the face may also be involved in tolerance of occlusion. In particular, research in the domain of object recognition has shown that observers have the capacity to perceptually "complete" objects when they are only partially specified due to occlusion (e.g., Sekuler, 1994; Sekuler, Palmer, & Flynn, 1994). Moreover, this process may be facilitated by object symmetry (e.g., Vetter, Poggio, & Bülthoff, 1994; Yokota, 1994) and has been observed for the perception of faces (Nakayama, Shimojo, & Silverman, 1989; see also Braddick, 1988). Consequently, perception of an occluded talking face may benefit particularly from the symmetry that exists about the vertical meridian, where featural arrangements (eyes, cheeks, etc.) and the distortion of features important for speech production (e.g., vertical mouth opening, horizontal lip stretching, jaw movement) occur to similar extents on both sides of the face.

However, the human face is not perfectly symmetrical (e.g., Bennett, Delmonico, & Bond, 1987; Carbary, Almierigi, & Harris, 1999; Gilbert & Bakan, 1973; Jordan & Thomas, 2007; Lindzey, Prince, & Wright, 1952; McCurdy, 1949; Seinen & Van der Werff, 1969; Wolff, 1933). Indeed, research measuring facial movements has revealed evidence of hemiface asymmetries in talkers' articulations, where the left side of the mouth (we refer to locations on the face as an observer sees them; e.g., left = left from an observer's point of view) opens sooner and wider during speaking, probably due to left hemisphere control over speech production (Graves, 1983; Graves, Goodglass, & Landis, 1982; Wolff, 1933; Wolf & Goodale, 1987). Moreover, this asymmetry in speech production may affect visual and audiovisual speech perception. Campbell (1986; see also Campbell, 1982) presented chimeric static faces formed from the left and right hemifaces of a speaker

articulating different speech sounds, such that one half of each chimeric face articulated one sound while the other half articulated another. When asked to identify the sounds spoken by the two halves of each chimeric face, participants reported the sound featured on the left hemiface of the speaker more accurately than that on the right, suggesting that the left hemiface is normally more important than the right for visual speech perception. Comparable findings have been obtained by Nicholls, Searle, and Bradshaw (2004), who found that when only the left or right half of a speaker's mouth (and no other part of the face) was occluded, the McGurk effect was weaker with left-half occlusions. Consequently, if symmetry helps ameliorate the effects of simple substantial occlusion on visual and audiovisual speech perception, previous evidence suggests that this amelioration may be less when occlusions cover the left side of a speaker's face.

Accordingly, the purpose of the present study was to investigate the tolerance of occlusion shown by visual and audiovisual speech perception using a range of simple substantial occlusions, each occluding an extensive area of the facial image by using an intervening opaque layer that gave the appearance of the face behind an occluding surface. Experiment 1 took the first step by using vertical occlusions in which an entire hemiface of a talking face (either the left or the right; see Fig. 1) was covered. This enabled effects of simple substantial occlusion on perception of visual and audiovisual speech to be investigated under conditions well-suited to benefits of vertical facial symmetry and to revealing the relative influences of left and right hemifaces when faces were substantially occluded. If either of the hemifacial occlusions used in the experiment impaired perception of visual speech, visual speech should be perceived more accurately and have a greater effect on auditory speech perception when presented in an unoccluded talking face. Moreover, if the effects observed with hemifacial occlusions support the effects of facial asymmetry reported previously (Campbell, 1982, 1986; Nicholls et al., 2004), visual speech should be perceived more accurately and have a greater effect on auditory speech perception in right-occluded displays than in left-occluded displays, because right-occluded displays reveal the speaker's left hemiface.

Experiment 1

Method

Participants A total of 18 native speakers of British English, 18–25 years of age, participated in the experiment. All participants had normal or corrected-to-normal visual acuity, as determined by a Bailey–Lovie eye chart, and were

Fig. 1 Facial displays used in Experiment 1: Unoccluded, left occluded, and right occluded

Unoccluded

Left-occluded

Right-occluded

right-handed, as determined by a revised Annett Handedness Questionnaire (Annett, 1970).

Stimuli and apparatus Four types of speech stimulus were used: auditory (in which auditory speech was presented with no visual speech), visual (in which visual speech was presented with no auditory speech), congruent audiovisual (in which visual and auditory speech were matched, as in normal speech), incongruent audiovisual (in which visual and auditory speech were not matched, as in the McGurk effect). In order to clarify the effects of simple substantial occlusion and to deemphasize effects of redundancy provided by scriptual, expressive, and contextual constraints (see, e.g., Gailey, 1987; Ronneberg, 1993; Samuelsson & Ronneberg, 1993; Sudman & Berger, 1971), the spoken stimuli were individual syllables articulated by an expressionless facial image.

Stimuli were created by recording the face of a 25-year-old male native English speaker while he fixated a video camera (JVC KY15). Frame-by-frame analysis of the recorded

footage confirmed that the speaker exhibited the faster and larger mouth movement on the left side of the face typical of speakers (e.g., Graves, 1983; Graves et al., 1982; Wolff, 1933; Wolf & Goodale, 1987). In particular, at the start of each articulation, the mouth moved 50 ms sooner on the left than on the right, and produced a vertical opening at the peak of each articulation that was 3 mm wider on the left than on the right. To remove any confounding influences of occlusion by head and facial hair, the speaker had short head hair and was clean shaven. The speaker's face was fully illuminated and recorded with his head stationary against a uniform dark-gray background with only the face and upper neck visible. The camera and lighting were configured so that the recorded face was reproduced life-sized and in natural color on a color display monitor that was used to monitor recordings and to display facial images in the experiment. Luminance was equated across the left and right hemifaces, as determined by a Minolta photometer at the speaker's face and a Cambridge Research Systems Color-CAL colorimeter at the monitor.

Recordings were made of the speaker saying each of six consonant–vowel utterances several times in a British English accent. Each utterance comprised a consonant—/b/, /g/, or /v/—plus one of two vowels—/a/ or /i/; these syllables are known to produce powerful effects of visual speech on auditory speech perception when presented with whole-face stimuli (e.g., Green, Kuhl, & Meltzoff, 1988; Jordan & Bevan, 1997; Jordan et al., 2000; Jordan & Thomas, 2001; Rosenblum et al., 2000). Each articulation began and ended with the face in neutral repose (mouth closed), and each syllable was spoken naturally with no artificial emphasis on articulation. These recordings were imported into digital postproduction editing software (Radius EditDV; Puffin Designs Commotion Pro; Adobe After Effects) running on a Macintosh G5. A single clip of each audiovisual syllable (/ba/, /bi/, /ga/, /gi/, /va/, or /vi/) was then selected, and the auditory and visual components of these clips were used to produce the experimental stimuli.

For auditory stimuli, the auditory component of each syllable was presented with a static image of the facial display (unoccluded, left occluded, or right occluded) in its neutral repose. This condition enabled the effects of facial movement to be dissociated from effects produced simply by the presence of a facial display, and so provided the most appropriate baseline control for assessing visual speech influences on auditory speech perception.[1] For visual stimuli, the visual component of each syllable (i.e., the face articulating) was presented with no auditory speech. For the congruent audiovisual

stimuli, each visual stimulus was presented with its auditory counterpart (auditory/visual: /ba/ba/, /bi/bi/, /ga/ga/, /gi/gi/, /va/va/, or /vi/vi/), as in normal speech. For incongruent audiovisual stimuli, each visual stimulus was presented with an incongruent auditory syllable to produce classic fusion (auditory /ba/ + visual /ga/ = usually heard /da/; auditory /bi/ + visual /gi/ = usually heard /di/), combination (auditory /ga/ + visual /ba/ = usually heard /bga/; auditory /gi/ + visual /bi/ = usually heard /bgi/), and visual-dominance (auditory /ba/ + visual /va/ = usually heard /va/; auditory /bi/ + visual /vi/ = usually heard /vi/) stimuli. The auditory, visual, congruent, and incongruent stimuli were formed from the same six recorded audiovisual clips, and all were matched for duration and auditory intensity.

Each stimulus was presented in three different displays: unoccluded, left occluded, and right occluded. Bisection of the face for occluded displays was determined using the Canonical Anthropometric Coordinate System (International Organization for Standardization, 2006). The landmarks used were the glabella (the midpoint between the eyebrows), the midpoint between the left and right endocanthion (the point at the inner commissure of each eye fissure), the labiale superius (the midpoint of the upper vermillion line of the upper lip), and the labiale inferius (the midpoint of the lower vermillion line of the lower lip). One hemiface (left or right) was then occluded by blacking out the image from the vertical midline of the face to the edges of the display screen (see Fig. 1). The blacked-out areas were darker than the background and gave the appearance of seeing a talking face behind a separate occluding surface.

The face was shown full size (20 cm from bottom of mandible to top of forehead) on the display monitor. Auditory signals were presented via two adjacent Spendor SP100 studio broadcast loudspeakers at a level of approximately 55 dB SPL over a background of continuous auditory white noise at a level of approximately 55 dB SPL. Pretesting had established that this signal-to-noise ratio produced approximately 70% correct responses to auditory stimuli presented without visual speech. This point on the performance scale was chosen to provide appropriate conditions for revealing effects of display type on the influences exerted by congruent and incongruent visual speech on auditory speech perception (see, e.g., Jordan & Sergeant, 1998; MacLeod & Summerfield, 1987, 1990; Middleweerd & Plomp, 1987; Sumby & Pollack, 1954). All sound measurements were made with a decibelometer at the location of each participant.

For the auditory stimuli, the face remained static throughout each display and the onset of auditory speech occurred 2 s after the onset of the face. For the visual, congruent audiovisual, and incongruent audiovisual stimuli, the face remained static until the onset of the appropriate visual or audiovisual articulation occurred, 2 s after the

---

[1] Preliminary testing showed that a blank screen produced the same pattern of performance in the auditory condition.

onset of the face. Each display lasted 6 s in total and was followed by a 5-s blank screen, during which participants made their response.

Design The stimuli were shown in three blocks corresponding to the three display types (unoccluded, left occluded, and right occluded). Each block comprised eight cycles of all 24 speech stimuli presented as a continuous sequence of 192 trials. Within each cycle, all auditory, visual, congruent audiovisual, and incongruent audiovisual speech stimuli were shown once, in a different random order for each cycle. Each participant saw all three blocks (unoccluded, left occluded, and right occluded) separated from each other by breaks of approximately 15 mins. The order in which the blocks were presented was counterbalanced across participants. Examples of the relevant display types were shown at the start of each block.

Procedure Each participant was seated in a sound-attenuated room at a table 1 m in front of the display screen, with his or her head level with the screen and supported on a chinrest. On the table was a sheet on which were printed 12 possible responses: "ba," "bi," "bga," "bgi," "da," "di," "ga," "gi," "tha," "thi," "va," and "vi." Pretesting had established that these responses constituted more than 97% of participants' perceptions of all stimuli used in the experiment. When questioned at the end of the experiment, all participants indicated that they had not been restricted in their responses. Participants were instructed to look at the screen, listen throughout each trial, and make their responses by marking on the response sheet the syllable they heard on each trial. For visual stimuli (i.e., stimuli without sound), participants were instructed to mark the syllable they saw articulated.

Results

The mean identification accuracy for each display type (unoccluded, left occluded, or right occluded) and speech condition (auditory, visual, congruent, or incongruent) is shown in Fig. 2. An ANOVA for identification accuracy was conducted for each speech condition.[2] Previous analyses for each speech condition had shown no main effects or interactions across block presentation orders or cycles (all Fs < 1).

---

[2] All analyses in this study were conducted on the raw data (proportions correct) and on arcsine transformations of these data. Both types of analysis revealed the same pattern of results and indicated no influence of floor or ceiling effects on performance. Consequently, for succinctness, we report only analyses of the raw data.
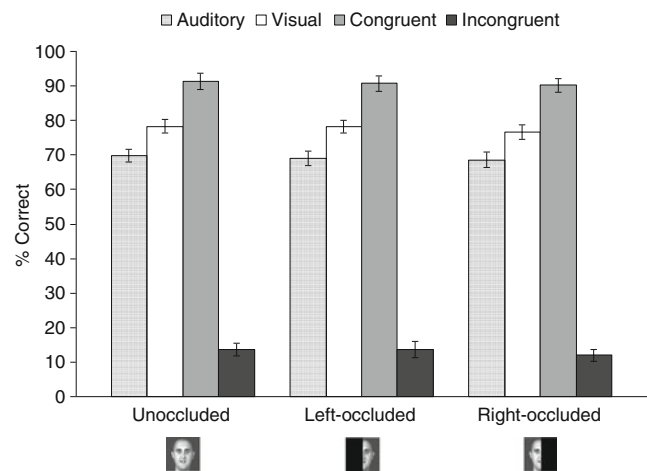
Fig. 2 Mean percentages and standard errors for auditory, visual, congruent, and incongruent stimuli correctly identified (% correct) in Experiment 1. All stimuli were presented in background auditory noise

Auditory An ANOVA with the factors Display Type and Speech Stimulus (/ba/, /bi/, /ga/, /gi/, /va/, /vi/) revealed a main effect of speech stimulus, $F(5, 85) = 30.10$, $MSE = 734.48$, $p < .001$, no main effect of display type, $F(2, 34) = 0.15$, $MSE = 265.82$, $p = .86$, and no interaction, $F(10, 170) = 1.02$, $MSE = 166.82$, $p = .43$. Newman–Keuls tests showed that /ba/ (50%) and /bi/ (52%) produced fewer correct responses than all other speech stimuli (/ga/, 77%; /gi/, 79%; /va/, 76%; /vi/, 78%; all ps < .01). Chi-square tests showed similar errors across all display types [all χ^2s(2) < 1.34, ps > .10]; errors for each of the following stimuli are in parentheses, with each stimulus set in bold for visibility: /**ba**/ (/da/, /ga/, /tha/, /va/); /**bi**/ (/di/, /thi/, /vi/); /**ga**/ (/ba/, /bga/, /da/); /**gi**/ (/bi/, /bgi/, /di/, /thi/, /vi/); /**va**/ (/ba/, /tha/); /**vi**/ (/bi/, /di/, /thi/).

Visual An ANOVA with the factors Display Type and Speech Stimulus (/ba/, /bi/, /ga/, /gi/, /va/, /vi/) revealed a main effect of speech stimulus, $F(5, 85) = 14.18$, $MSE = 902.25$, $p < .001$, no main effect of display type, $F(2, 34) = 2.36$, $MSE = 185.21$, $p = .20$, and no interaction, $F(10, 170) = 1.07$, $MSE = 170.91$, $p = .38$. Newman–Keuls tests showed that /ga/ (62%) and /gi/ (63%) produced fewer correct responses than all other speech stimuli (/ba/, 86%; /bi/, 87%; /va/, 85%; /vi/, 86%; all ps < .01). Chi-square tests showed similar errors across all display types [all χ^2s(2) < 2.92, ps > .10]: /**ba**/ (/bga/, /va/); /**bi**/ (/bgi/, /vi/); /**ga**/ (/bga/, /da/, /tha/); /**gi**/ (/bgi/, /di/, /thi/); /**va**/ (/ba/, /tha/); /**vi**/ (/thi/).

Congruent audiovisual An ANOVA with the factors Display Type and Speech Stimulus (auditory/visual: /ba/ba/, /bi/bi/, /ga/ga/, /gi/gi/, /va/va/, /vi/vi/) revealed a main effect of speech stimulus, $F(5, 85) = 3.61$, $MSE = 263.48$,

$p < .01$, no main effect of display type, $F(2, 34) = 1.26$, $MSE = 34.67$, $p = .29$, and no interaction, $F(10, 170) = 1.39$, $MSE = 35.62$, $p = .19$. Newman–Keuls tests showed that /ba/ba/ (83%) and /bi/bi/ (82%) produced fewer correct auditory responses than all other speech stimuli (/ga/ga/, 95%; /gi/gi/, 95%; /va/va/, 94%; /vi/vi/, 95%; all ps < .01). No other comparisons were significant. Chi-square tests showed similar errors across all display types [all χ^2s(2) < 3.18, ps > .10]: **/ba/ba/** (/bga/, /va/); **/bi/bi/** (/bgi/, /vi/); **/ga/ga/** (/bga/, /da/); **/gi/gi/** (/bgi/, /di/); **/va/va/** (/tha/); **/vi/vi/** (/thi/).

Further analysis showed that improvements in auditory perception produced by congruent visual speech relative to the auditory condition were also unaffected by occlusion. An ANOVA with the factors Speech Condition (auditory, congruent), Display Type, and Speech Stimulus showed a main effect of speech condition, $F(1, 17) = 75.25$, $MSE = 590.87$, $p < .001$, due to overall higher response accuracy in the congruent condition, and an interaction between speech condition and speech stimulus, $F(5, 85) = 23.93$, $MSE = 331.26$, $p < .001$. No evidence of a main effect or interaction involving display type was obtained (all Fs < 1.42). Newman–Keuls tests revealed that congruent visual speech produced more correct auditory responses for all speech stimuli (mean increase = 22%, ps < .01), although the increase was greatest for /ba/ba/ (33%) and /bi/bi/ (30%) (ps < .001), where /ba/ and /bi/ were least well identified in the auditory condition.

Incongruent audiovisual A response to incongruent audiovisual stimuli was correct when the auditory signal was identified correctly. An ANOVA with the factors Display Type and Speech Stimulus (auditory/visual: /ba/ga/, /bi/gi/, /ga/ba/, /gi/bi/, /ba/va/, /bi/vi/) revealed a main effect of speech stimulus, $F(5, 85) = 27.04$, $MSE = 685.47$, $p < .001$, no main effect of display type, $F(2, 34) = 0.22$, $MSE = 159.25$, $p = .81$, and no interaction, $F(10, 170) = 0.32$, $MSE = 114.82$, $p = .98$. Newman–Keuls tests showed that, for all display types, /ga/ba/ (29%) and /gi/bi/ (28%) produced more correct auditory responses (i.e., fewer McGurk effects) than any other speech stimulus (/ba/ga/, 7%; /bi/gi/, 8%; /ba/va/, 4%; /bi/vi/, 3%; all ps < .001). Chi-square tests showed similar errors across all display types [all χ^2s(2) < 2.68, ps > .10]: **/ba/ga/** (/da/, /tha/); **/bi/gi/** (/di/, /thi/); **/ga/ba/** (/bga/); **/gi/bi/** (/bgi/); **/ba/va/** (/tha/, /va/); **/bi/vi/** (thi/, /vi/).

Further analysis showed that impairments in auditory perception produced by incongruent visual speech relative to the auditory condition were also unaffected by occlusion. An ANOVA with the factors Speech Condition (auditory, incongruent), Display Type, and Speech Stimulus showed only a main effect of speech condition, $F(1, 17) = 463.60$, $MSE = 1,267.81$, $p < .001$, reflecting overall fewer correct auditory responses in the incongruent condition. No evidence of a main effect or interaction involving display type was obtained (all Fs < 1.12).

Discussion

The findings of Experiment 1 show that perception of visual and audiovisual speech was unaffected by simple substantial occlusion of facial information. In particular, both the identification of visual speech and the effects of visual speech on identification of auditory speech observed with unoccluded faces were unchanged when an entire hemiface was occluded. Moreover, no differences in performance were observed between left-occluded and right-occluded displays, despite previous suggestions that left hemifaces convey more information than right ones (e.g., Campbell, 1982, 1986; Nicholls et al., 2004). Indeed, these findings show that the similarities in performance observed across types of display are reflected not only by the levels of accurate speech identification obtained, but also by the patterns of errors that were produced by each speech stimulus. The reasonable implication is, therefore, that the same information underlying perception of visual speech and its influence on auditory speech perception in unoccluded facial displays can be obtained even when only one hemiface (left or right) is visible. Moreover, performance was unaffected by the order in which each occlusion condition (block) was presented or the point in each block (cycle) when stimuli were shown. This suggests that the findings obtained were not the result of learning the speech stimuli or occlusion displays used in the experiment, but instead reflect natural tolerance of simple substantial occlusion in visual and audiovisual speech perception.

Experiment 2 investigated the effects of simple substantial occlusion on visual and audiovisual speech perception further by using horizontal and diagonal occlusions (Fig. 3). These displays maintained the characteristics of simple substantial occlusion that are the focus of this study and permitted similar-sized areas of the face to be occluded across conditions. Left-occluded and right-occluded displays were included in Experiment 2 to replicate with new participants the tolerance of occlusion observed in Experiment 1 and to provide within-experiment comparisons with the other occlusions. In addition, to further test the robustness of this tolerance of occlusion, the blocked design of Experiment 1 was replaced by a randomly interleaved design in which the occlusion on each trial was unpredictable. Thus, Experiment 2 investigated the influence of nine different facial occlusions (unoccluded, left occluded, right occluded, upper occluded, upper left occluded, upper right occluded, lower occluded, lower left occluded, and

lower right occluded) on perception of visual and audiovisual speech (Fig. 3).
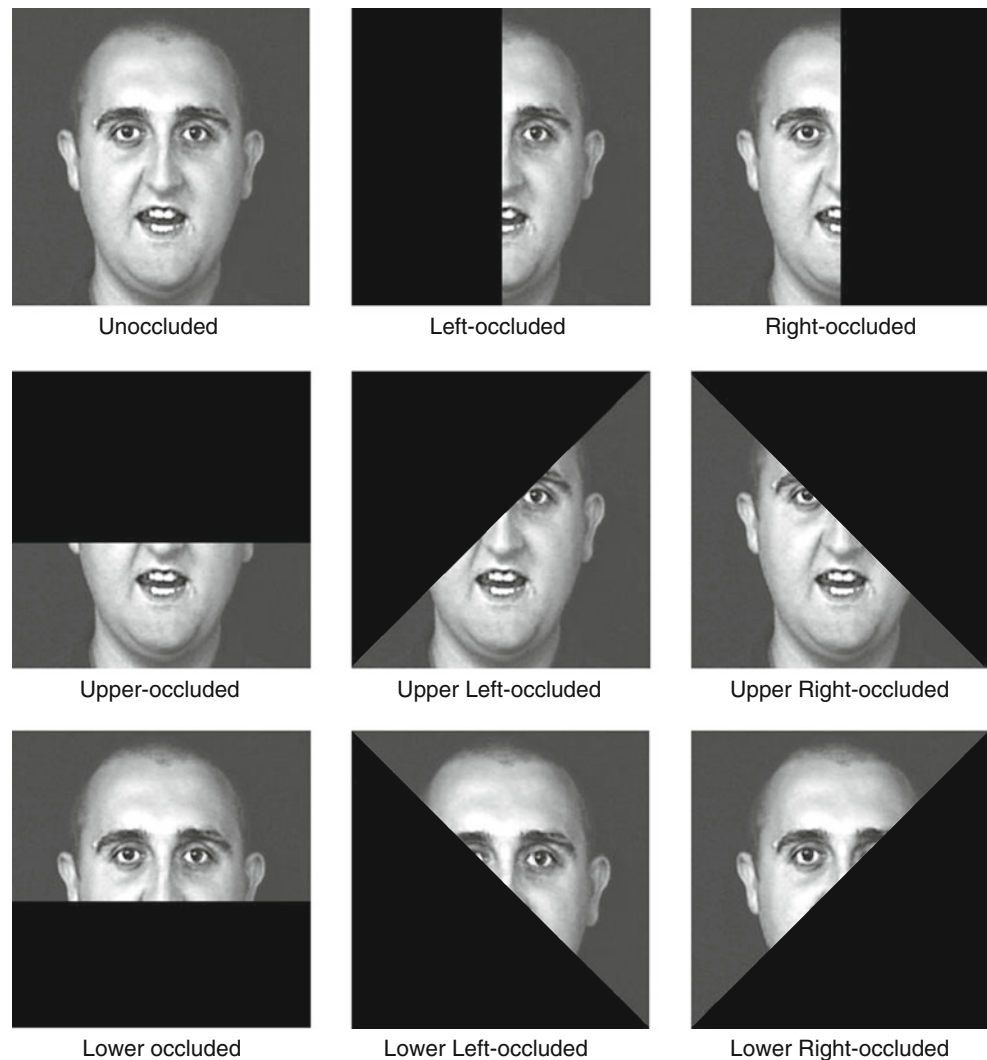
Experiment 2

Method

Participants A total of 16 native speakers of British English, 18–25 years of age, from the same population as Experiment 1 participated in the experiment. All participants had normal or corrected-to-normal visual acuity, as determined by a Bailey–Lovie eye chart, and were right-handed, as determined by a revised Annett Handedness Questionnaire (Annett, 1970). None had participated in Experiment 1.

Stimuli and design The same auditory, visual, congruent, and incongruent stimuli used to produce the displays in Experiment 1 were used to produce the displays in Experiment 2. Each stimulus was presented in an unoccluded display and in eight occluded displays. Left-occluded and right-occluded displays were constructed as in Experiment 1. The horizontal occluding edges of the upper occluded and lower occluded displays passed through a midpoint located on the vertical median (defined as in Exp. 1) halfway between the uppermost and lowermost points of the face (see Fig. 3). These occlusions were then rotated 45º to produce displays that were occluded diagonally (upper left occluded, upper right occluded, lower left occluded, and lower right occluded).

Stimuli were shown in eight blocks of 216 trials. Within each block, each auditory, visual, congruent, and incongruent speech stimulus was shown once in all nine display types (unoccluded, left occluded, right occluded, upper occluded, upper left occluded, upper right occluded, lower occluded, lower left occluded, and lower right occluded). Examples of each display type were shown at the start of each block. Stimulus presentation in each block was randomly ordered, and blocks were separated by 15-min

Fig. 3 Facial displays used in Experiment 2: Unoccluded, left occluded, right occluded, upper occluded, upper left occluded, upper right occluded, lower occluded, lower left occluded, and lower right occluded

breaks. All other aspects were identical to those of Experiment 1.

Results

The mean identification accuracy for each display type (unoccluded, left occluded, right occluded, upper occluded, upper left occluded, upper right occluded, lower occluded, lower left occluded, or lower right occluded) and speech condition (auditory, visual, congruent, or incongruent) is shown in Fig. 4. An ANOVA for identification accuracy was conducted for each speech condition. Previous analyses for each speech condition showed no main effects or interactions across blocks (Fs < 1).

Auditory An ANOVA with the factors Display Type and Speech Stimulus revealed a main effect of speech stimulus, $F(5, 75) = 11.71$, $MSE = 3,295.86$, $p < .001$, no main effect of display type, $F(8, 120) = 1.33$, $MSE = 315.48$, $p < .24$, and no interaction, $F(40, 600) = 1.11$, $MSE = 279.24$, $p < .30$. Newman–Keuls tests showed that /ba/ (56%) and /bi/ (58%) produced fewer correct responses than all other speech stimuli (/ga/, 78%; /gi/, 76%; /va/, 78%; /vi/, 79%; all ps < .01). Chi-square tests showed similar errors across all display types [all χ^2s(2) < 1.61, ps > .10]; again, errors are shown in parentheses: /**ba**/ (/da/, /ga/, /tha/, /va/); /**bi**/ (/di/, /thi/, /vi/); /**ga**/ (/ba/, /bga/, /da/); /**gi**/ (/bi/, /bgi/, /di/, /thi/, /vi/); /**va**/ (/ba/, /tha/); /**vi**/ (/bi/, /di/, /thi/).

Visual An ANOVA with the factors Display Type and Speech Stimulus revealed main effects of display type,

$F(8, 120) = 230.40$, $MSE = 357.31$, $p < .001$, and speech stimulus, $F(5, 75) = 7.18$, $MSE = 1,347.42$, $p < .001$, and an interaction, $F(40, 600) = 4.33$, $MSE = 345.23$, $p < .001$. Newman–Keuls tests showed no differences in overall response accuracy across unoccluded (79%), left-occluded (80%), right-occluded (80%), upper-occluded (79%), upper-left occluded (81%), and upper-right occluded (81%) displays, or across lower-occluded (19%), lower-left occluded (20%), and lower-right occluded (20%) displays. However, lower-occluded, lower-left occluded, and lower-right occluded displays produced fewer correct responses than all other displays (ps < .001). These decreases in response accuracy were observed for all stimuli (all ps < .01) but were less for /ga/ (mean decrease 39%) and /gi/ (mean decrease 42%) than for /ba/ (mean decrease 69%), /bi/ (mean decrease 72%), /va/ (mean decrease 68%), and /vi/ (mean decrease 69%) (all ps < .01). For unoccluded, left-occluded, right-occluded, upper-occluded, upper-left occluded, and upper-right occluded displays, /ga/ (mean 60%) and /gi/ (mean 63%) produced fewer correct responses than all other stimuli (/ba/, 89%; /bi/, 90%; /va/, 88%; /vi/ 90%; all ps < .01). For lower-occluded, lower-left occluded, and lower-right occluded displays, all stimuli produced similar levels of performance (means: /ba/, 20%; /bi/, 18%; /ga/, 21%; /gi/, 21%; /va/, 20%; /vi/, 21%), and all were substantially above chance (chance = 8.33%, all ps < .01).

Chi-square tests showed that unoccluded, left-occluded, right-occluded, upper-occluded, upper-left occluded, and upper-right occluded displays produced similar errors [all χ^2s(2) < 2.64, ps > .10]: /**ba**/ (/bga/, /va/); /**bi**/ (/bgi/); /**ga**/ (/bga/, /da/, /tha/); /**gi**/ (/bgi/, /di/, /thi/); /**va**/ (/ba/, /tha/); /**vi**/
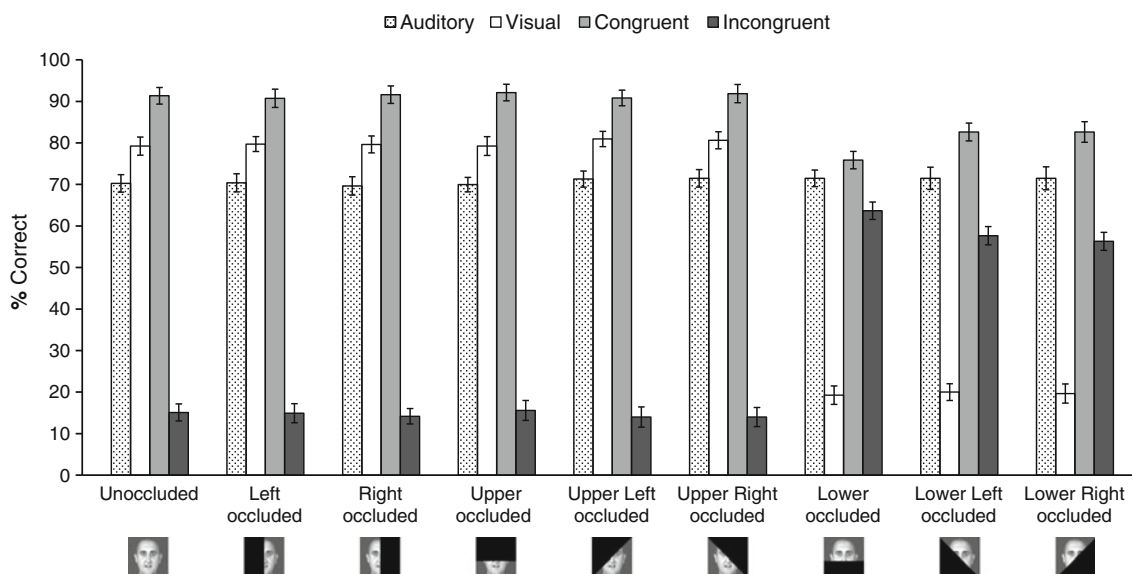

Fig. 4 Mean percentages and standard errors for auditory, visual, congruent, and incongruent stimuli correctly identified (% correct) in Experiment 2. All stimuli were presented in background auditory noise

(/thi/). Lower-occluded, lower-left occluded, and lower-right occluded displays showed a wider spread of errors that was shared across these three display types [all χ^2s(2) < 1.25, ps > .10]: /**ba**/ (/bga/, /da/, /ga/, /tha/, /va/); /**bi**/ (/bgi/, /di/, /gi/, /thi/, /vi/); /**ga**/ (/ba/, /bga/, /da/, /tha/, /va/); /**gi**/ (/bi/, /bgi/, /di/, /thi/, /vi/); /**va**/ (/ba/, /bga/, /da/, /ga/, /tha/); /**vi**/ (/bi/, /bgi/, /di/, /gi/, /thi/).

Congruent audiovisual An ANOVA with the factors Display Type and Speech Stimulus revealed main effects of display type, $F(8, 120) = 33.38$, $MSE = 176.43$, $p < .001$, and speech stimulus, $F(5, 75) = 10.14$, $MSE = 952.12$, $p < .001$, and an interaction, $F(40, 600) = 11.00$, $MSE = 154.41$, $p < .001$. Newman–Keuls tests showed no differences in overall levels of auditory response accuracy across unoccluded (91%), left-occluded (91%), right-occluded (92%), upper-occluded (92%), upper-left occluded (91%), and upper-right occluded (92%) displays, or across lower-left (83%) and lower-right (83%) occluded displays. However, lower-occluded, lower-left occluded, and lower-right occluded displays produced fewer correct responses than all other displays (all ps < .001), and lower-occluded displays (76%) produced the fewest correct responses of all (ps < .01). These decreases in auditory response accuracy for lower-occluded, lower-left occluded, and lower-right occluded displays occurred for all stimuli (ps < .01), but the decreases were less for /ga/ga/ (mean decrease 5%) and /gi/gi/ (mean decrease 6%) than for /ba/ba/ (mean decrease 15%), /bi/bi/ (mean decrease 17%), /va/va/ (mean decrease 14%), and /vi/vi/ (mean decrease 12%) (ps < .01). For unoccluded, left-occluded, right-occluded, upper-occluded, upper-left occluded, and upper-right occluded displays, /ba/ba/ (mean 81%) and /bi/bi/ (mean 82%) produced fewer correct auditory responses (ps < .01) than all other stimuli (means: /ga/ga/, 99%; /gi/gi/, 99%; /va/va/, 95%; /vi/vi/, 94%). For lower-occluded, lower-left occluded, and lower-right occluded displays, /ba/ba/ (mean 66%), /bi/bi/ (mean 65%), /va/va/ (mean 81%), and /vi/vi/ (mean 82%) produced the fewest correct auditory responses, and /ga/ga/ (mean 94%) and /gi/gi/ (mean 93%) produced the most (ps < .01).

Chi-square tests showed that, although more errors were made with lower-occluded, lower-left occluded, and lower-right occluded displays, similar errors were observed across all display types [all χ^2s(2) < 2.12, ps > .10]: /**ba/ba**/ (/bga/, /tha/, /va/); /**bi/bi**/ (/bgi/, /thi/, /vi/); /**ga/ga**/ (/bga/, /da/, /tha/); /**gi/gi**/ (/bgi/, /di/, /thi/); /**va/va**/ (/ba/, /tha/); /**vi/vi**/ (/bi/, /thi/).

An ANOVA with the factors Speech Condition (auditory, congruent audiovisual), Display Type, and Speech Stimulus revealed a main effect of speech condition, $F(1, 15) = 67.82$, $MSE = 1,505.41$, $p < .001$, due to overall higher accuracy in the congruent condition. There were also interactions between speech condition and display type, $F(8, 120) = 16.08$, $MSE = 227.93$, $p < .001$, speech condition and speech

stimulus, $F(5, 75) = 6.96$, $MSE = 1,042.61$, $p < .001$, and speech condition, display type, and speech stimulus, $F(40, 600) = 5.40$, $MSE = 201.04$, $p < .001$. For /ga/ga/ and /gi/gi/, congruent visual speech produced significant increases in auditory response accuracy (ps < .01) for all display types (mean increase 19%), including lower-occluded (mean increase 16%), lower-left occluded (mean increase 17%), and lower-right occluded (mean increase 17%) displays. For all other stimuli, all display types except lower occluded ($p > .80$) produced significant increases in auditory response accuracy (mean increase 18%, all ps < .01), although these increases were less for lower-left occluded (mean 8%) and lower-right occluded (mean 9%) displays (ps < .01).

Incongruent audiovisual A response to incongruent audiovisual stimuli was correct when the auditory signal was identified correctly. An ANOVA with the factors Display Type and Speech Stimulus revealed main effects of display type, $F(8, 120) = 136.38$, $MSE = 481.18$, $p < .001$, and speech stimulus, $F(5, 75) = 17.40$, $MSE = 1,886.26$, $p < .001$, and an interaction, $F(40, 600) = 5.90$, $MSE = 286.03$, $p < .001$. Newman–Keuls tests showed no differences in overall levels of auditory response accuracy across unoccluded (15%), left-occluded (15%), right-occluded (14%), upper-occluded (16%), upper-left occluded (14%), and upper-right occluded (14%) displays, or across lower-left occluded (58%) and lower-right occluded (56%) displays. However, lower-occluded, lower-left occluded, and lower-right occluded displays produced more correct responses (i.e., fewer McGurk effects) than all other displays (ps < .001), and lower-occluded displays (64%) produced the most correct responses of all (ps < .01). These increases in auditory response accuracy for lower-occluded, lower-left occluded, and lower-right occluded displays occurred for all stimuli (ps < .01), but the increases were less for /ba/ga/ (mean increase 40%) and /bi/gi/ (mean increase 39%) than for /ba/va/ (mean increase 47%), /bi/vi/ (mean increase 49%), /ga/ba/ (mean increase 47%), and /gi/bi/ (mean increase 45%) (all ps < .01). For unoccluded, left-occluded, right-occluded, upper-occluded, upper-left occluded, and upper-right occluded displays, /ba/va/ (mean 5%) and /bi/vi/ (mean 4%) produced fewer correct auditory responses than all other stimuli (ps < .01), and /ba/ga/ (mean 11%) and /bi/gi/ (mean 12%) produced fewer correct auditory responses than /ga/ba/ (mean 27%) and /gi/bi/ (mean 29%) (ps < .01). For lower-occluded, lower-left occluded, and lower-right occluded displays, /ba/ga/ (mean 50%), /bi/gi/ (mean 51%), /ba/va/ (mean 52%), and /bi/vi/ (mean 53%) produced the fewest correct auditory responses, and /ga/ba/ (mean 74%) and /gi/bi/ (mean 75%) produced the most (ps < .01).

Chi-square tests showed that, although fewer errors were made with lower-occluded, lower-left occluded, and lower-right occluded displays, similar errors were observed

across all display types [all χ^2s(2) < 2.03, ps > .10]: /**ba**/**ga**/ (/da/, /tha/, /va/); /**bi**/**gi**/ (/di/, /thi/, /vi/); /**ga**/**ba**/ (/ba/, /bga/, /da/); /**gi**/**bi**/ (/bi/, /bgi/, /di/); /**ba**/**va**/ (/tha/, /va/); /**bi**/**vi**/ (/thi/, /vi/).

An ANOVA with the factors Speech Condition (auditory, incongruent), Display Type, and Speech Stimulus revealed a main effect of speech condition, $F(1, 15) = 676.65$, $MSE = 1,267.15$, $p < .001$, due to overall lower accuracy in the incongruent condition. There were also interactions between speech condition and display type, $F(8, 120) = 78.74$, $MSE = 411.59$, $p < .001$, speech condition and speech stimulus, $F(5, 75) = 5.51$, $MSE = 3,170.45$, $p < .001$, and speech condition, display type, and speech stimulus, $F(40, 600) = 3.13$, $MSE = 270.69$, $p < .001$. For /ba/ga/ and /bi/gi/, incongruent visual speech produced significant decreases (ps < .01) in auditory response accuracy for all display types (mean decrease 34%), including lower-occluded (mean decrease 6%), lower-left occluded (mean decrease 7%), and lower-right occluded (mean decrease 6%) displays. For all other stimuli, all display types except lower occluded ($p > .70$) produced significant decreases (mean 42%, ps < .001), although the decreases were much less for lower-left occluded (mean 7%) and lower-right occluded (mean 9%) displays (ps < .01).

Discussion

Experiment 2 replicated the finding of Experiment 1 that the identification of visual speech and the effects of visual speech on identification of auditory speech obtained with unoccluded faces remained unchanged when either hemiface, left or right, was completely occluded. Moreover, this replication was observed despite substantial increase in the variability of facial occlusions in Experiment 2 and the change from a blocked to a randomly interleaved design, indicating the robustness of the tolerance of occlusion observed in these experiments. Experiment 2 also demonstrated considerable tolerance of nonvertical occlusions. In particular, displays in which upper areas of the face were occluded (upper occluded, upper left occluded, and upper right occluded) produced levels of visual and audiovisual speech performance that were equal to those obtained with unoccluded faces, and these similarities in performance across display types were reflected not only by levels of stimulus identification but also by the patterns of errors produced. Displays in which lower areas of the face were occluded (lower occluded, lower left occluded, and lower right occluded) also showed evidence of visual and audiovisual speech perception, although these displays produced generally lower levels of visual speech perception and less influence on auditory speech perception than other displays. Indeed, identification of visual speech was

particularly impaired when lower areas of the face were occluded (lower occluded, lower left occluded, and lower right occluded), dropping to an average of approximately 20% from an average of approximately 80% across the six other display types. Nevertheless, even with these lower facial occlusions, identification of visual speech was considerably above chance, vowel identification was accurate, and congruent and incongruent audiovisual conditions showed significant influences of visual speech on auditory speech perception (although the effects were more restricted for lower-occluded displays) and produced patterns of errors similar to those produced by unoccluded displays. Moreover, the greater effects of occlusion on visual and audiovisual speech perception in lower-occluded, lower-left occluded, and lower-right occluded displays were most apparent for stimuli containing /ba/, /bi/, /va/, and /vi/ visual speech (unimodal, congruent, and incongruent), suggesting that, with these lower facial occlusions, visual processing was disrupted most for these articulations and less for /ga/ and /gi/. We return to these points in the General Discussion.

General discussion

The experiments reported in this article investigated effects of simple substantial occlusions on the perception of visual and audiovisual speech. The motivations for this approach were the likelihood that visual and audiovisual speech perception generally copes with naturally occurring occlusions that are not so specific as to hide all of a face except for a particular feature, as well as the concern that experimental occlusions that reveal only individual features may inflate a feature's influence and distort understanding of the influence of other areas of the face. Consequently, simple substantial occlusions should help provide a more complete indication of the tolerance of visual and audiovisual speech perception to occlusion. Over both experiments, vertically, horizontally, and diagonally occluded facial displays revealed widespread tolerance of occlusion. Some vulnerability was observed with displays in which lower areas of the face were occluded (lower occluded, lower left occluded, and lower right occluded). Nevertheless, even with these displays, visual speech perception and visual speech influences on auditory speech perception were still evident.

These findings provide new indications that visual speech information is distributed widely across the facial surface, with the effect that a range of facial areas provides sources of visual speech information that help offset the potentially disruptive effects of substantial occlusion. Perhaps the most straightforward explanation of this tolerance of occlusion would be that, when an occluded

face is encountered, the visible area activates an independently stored perceptual representation for that area, and this activation provides sufficient self-contained information for visual speech perception to take place. In this way, a visible facial area (e.g., one hemiface or the lower part of a face) may provide sufficient information for visual speech perception equal to that obtained with unoccluded faces, and other areas (e.g., the upper part of a face) may provide information sufficient for at least some visual speech perception, although less than that obtained with unoccluded faces. For this purely "bottom-up" processing to function, representations must exist for just those areas of a face that are visible, and separate representations would presumably develop for the different areas of a face that are visible sufficiently often in encounters with occluded faces, and each representation would reflect the informativeness of just that area. So, for example, when just the left hemiface is visible in an otherwise occluded image, a previously stored representation of just the left hemiface would be activated, independently of information from the remainder of the face.

However, although environmental occlusion of faces is common, faces are obscured in various uncontrolled ways that present the observer with a range of occluded views of the same broad facial area (e.g., occlusion of the right hemiface) but not views that correspond consistently with the precise location and extent of a particular occlusion (e.g., the precise occlusion of all the right hemiface, no less and no more). Consequently, it is unclear that the precise and specific occlusion of a particular facial area (e.g., the precise occlusion of all the right hemiface) would generally be encountered with sufficient frequency and regularity for individual representations for such specifically occluded views to develop in an observer. Moreover, while the occlusions used in our experiments revealed effects of simple substantial occlusion, it seems unlikely that participants had previously experienced the images that were used, and so had not previously developed individual representations for the precise occlusions that were displayed.

Consequently, it seems unlikely that the effects observed in our experiments reflect activation of representations corresponding to the precise location and extent of particular occlusions (and analyses indicate that this situation was not altered by repeated exposure to the same occlusions within each experiment). However, bottom-up processes of visual speech perception may involve access to more complete representations of talking facial information, such that similar levels of activation in these representations are achieved by a range of occluded views, even occluded views that are not identical to any previously encountered. In this way, visual speech may be perceived via representations specifying the various sources of visual speech across

the face and the relations between these sources represented by structural descriptions (e.g., Biederman & Gerhardstein, 1993, 1995; Lowe, 1987). As long as important sources of visual speech are visible, the same structural description will be activated by a range of images in which different parts of the face are occluded. For example, the visual invariants of the size and shape of the mouth opening, and the rate of change of this information over time, may provide a major cue to visual speech (e.g., Jordan & Sergeant, 1998, 2000; Massaro & Cohen, 1996; Montgomery & Jackson, 1983; Munhall & Vatikiotis-Bateson, 1998; Petajan, 1984; Summerfield, 1987, 1992; Summerfield & McGrath, 1984), and this information may be available from many occluded views, including all but the lower occluded, lower-left occluded, and lower-right occluded displays used in our experiments. This would help explain why occlusions that revealed very different areas of the face (e.g., left occluded, right occluded, upper occluded, upper left occluded, and upper right occluded) produced near-identical performance, because the same structural description would be activated in each case. Moreover, since movements of other facial areas, such as the side of the jaw and the cheeks, are highly correlated with movements of the mouth (e.g., Munhall & Vatikiotis-Bateson, 1998), perception of these sources may be sufficient to allow for some access to the same structural description, even when visibility of the actual mouth is impaired by occlusion (as in lower-occluded, lower-left occluded, and lower-right occluded displays). Thus, the invariance of these relations across the face may be a major contribution to the formation of structural descriptions of visual speech that are not tied to specific facial views and help observers tolerate effects of substantial occlusion in the environment.

The process of activating representations of visual and audiovisual speech when faces are substantially occluded may involve other processes beyond the facial information that is visible. As described in the introductory section, symmetry and perceptual completion generally play a major role in object recognition (e.g., Vetter et al., 1994; Yokota, 1994), and the vertical symmetry that exists in talking faces may support a similar process for visual and audiovisual speech perception. In particular, because the featural arrangements (eyes, cheeks, etc.) and distortion of features important for visual speech perception (e.g., vertical and horizontal mouth opening, movement of the cheeks and jaw) occur on both sides of a face, these areas would often be duplicated behind an occluding surface. Consequently, when observing a talking face that is substantially occluded in the environment, representations for visual speech may often be activated both by the area of the face that is visible and by the area of the face that is not visible but that can be inferred from vertical symmetry. This additional source of activation through symmetry would help explain the

complete tolerance shown by visual and audiovisual speech perception when only the left or right hemiface was visible (i.e., in right-occluded and left-occluded displays), and might also have contributed to the perception of visual and audiovisual speech in diagonally occluded displays, where visible areas of the face were duplicated in areas hidden by occlusion (e.g., areas of cheek and jaw).

Occlusions that revealed lower areas of the face (including the oral area and jaw) produced levels of visual and audiovisual speech perception that were superior to those obtained when these areas were occluded. These findings are consistent with the effects of these individual areas when shown in isolation (e.g., Berger et al., 1971; Greenberg & Bode, 1968; IJsseldijk, 1992; Marassa & Lansing, 1995; Stone, 1957), and many theories of visual speech perception emphasize the importance of the oral area and the oral cavity (for a review, see Thomas & Jordan, 2004). However, the indication from both experiments reported here is that although left-occluded and right-occluded displays completely occluded half of the oral area, the visual and audiovisual speech perception observed with unoccluded faces was unaffected by this disruption. Thus, although the oral area is clearly important for visual and audiovisual speech perception, the potentially disruptive effect of occluding half of this information when completely occluding an entire hemiface can apparently be overcome. Moreover, the similar performance produced by left-occluded and right-occluded displays indicates that this remarkable tolerance to occlusion is independent of asymmetry in articulatory movement. This suggests that perception of visual speech in occluded facial displays reflects the importance of the invariant properties of basic visual cues to facial movement in visual speech, rather than surface detail such as the subtle differences that exist in the onset, rate, and extent of facial movement between the two hemifaces.

The absence of disruption to visual and audiovisual speech perception when an entire hemiface (left or right) was occluded does not support the view that the left hemiface conveys more information than the right (Campbell, 1982, 1986; Nicholls et al., 2004). This contrast with previous findings indicates important differences in the tolerance shown by visual and audiovisual speech perception towards different visual disruptions. Campbell (1982, 1986) used static chimeric faces formed from the left and right hemifaces of a speaker, such that one half of each chimeric face showed one sound, while the other half showed another. In the present study, each image was naturally dynamic and contained no contradictory hemiface information, suggesting that the natural effectiveness of the two hemifaces for visual and audiovisual speech perception may actually be identical when faces are viewed normally (but see Jordan & Thomas,

2007). Moreover, in the study by Nicholls et al. (2004), perception of information from the oral area was prevented by using a small patch to occlude just one half of the mouth (left or right) and no other part of the face. As we have described, visual and audiovisual speech perception is likely to develop to cope with everyday situations that leave the observer with a view of a talker's face in which an intervening surface produces a substantial and continuous area of occlusion. Consequently, a small patch that covers just half of the mouth and no other part of the face may present a rather unnatural set of circumstances that are difficult to accommodate, and so are particularly disruptive. In addition, a patch that covers just half of the mouth is likely to induce abnormal focus of visual and attentional resources on the oral area, which then accentuates perception of left–right differences in oral movement. Consequently, when part of the mouth only is occluded, normal visual speech perception may suffer, and the influence of left–right facial differences on visual speech perception may be exaggerated.[3]

Our findings also show that effects of occlusion can vary with the informativeness of facial locations for particular articulations. In Experiment 2, the occlusion of information in lower facial areas (lower occluded, lower-left occluded, lower-right occluded) showed evidence of affecting visual encoding of /ba/, /bi/, /va/, and /vi/ more than /ga/ and /gi/. Articulation of /ba/ and /bi/ involves a bilabial closure followed by an open mouth, and articulation of /va/ and /vi/ involves a labiodental closure followed by an open mouth, all of which are prominent and highly visible maneuvers when lower facial areas are unoccluded. In contrast, articulation of /ga/ and /gi/ involves less prominent visible movement (essentially, an open mouth throughout each articulation), and crucial articulatory maneuvers within the vocal tract (e.g., velar, alveolar) that are not normally

---

[3] The temptation to invoke hemispheric projections to explain differences in hemiface processing for facial displays deserves comment (see also Jordan & Thomas, 2007). First, numerous studies have indicated that there is an overlap at the midline of up to 3º where visual information will project to both hemispheres (for recent reviews, see Jordan & Paterson, 2009, 2010). Consequently, the left and right hemifaces of facial displays cannot be projected completely to separate hemispheres. This problem is compounded by the common practice of using facial stimuli that barely extend beyond the area of bilateral projection, making it likely that most of each face projects to both hemispheres, and this would usually be the case in face-to-face interactions in everyday life. Second, notwithstanding the issue of bilateral projections, participants would need to fixate a face along its midline for information in each hemiface to project accurately to the contralateral hemisphere. However, central fixation in freely viewed images cannot be ensured, and the findings of Butler et al. (2005) suggest that fixations of faces usually fall to the left of center (i.e., on the left hemiface), which would result in most of the face projecting to the left hemisphere. In short, the notion that faces provide stimuli in which information from left and right hemifaces are projected to different contralateral hemispheres should be treated with caution.

visible. Indeed, previous research (e.g., Walden, Prosek, Montgomery, Scherr, & Jones, 1977) has indicated that information from the lower facial area is highly informative when identifying bilabial and labiodental articulations like /b/ and /v/, but much less informative for identifying /g/. Thus, while the findings of the present study indicate that visual speech is encoded from a range of facial locations, it appears that occluding the lower facial area impairs visual speech perception most when this area contains the most visible articulations (see also Scheinberg, 1980). Specifically, when articulatory maneuvers are normally visible in the lower facial area (e.g., for /ba/, /bi/, /va/, /vi/), occluding this information is highly disruptive. In contrast, when articulatory maneuvers are not normally visible in the lower facial area (e.g., within the vocal tract for /ga/ and /gi/), considerable reliance may be placed on visible information from other, visible facial areas (including the upper part of the face), and the influence of these areas becomes particularly clear when the lower facial area is occluded.

It seems, therefore, that natural, repeated and prolonged exposure to the facial movements that accompany the production of speech sounds may produce high levels of familiarity with the patterns of movement that occur across different regions of the face, and this familiarity may help provide general stability for processing visual and audiovisual speech when perception of the oral area is not possible through occlusion. However, cues outside the oral area may be more subtle than cues within the oral area (see also Thomas & Jordan, 2004). Thus, while perception of some articulations may benefit particularly from extraoral facial movement (e.g., /ga/ and /gi/), effects of extraoral cues may often be hidden when the more salient cues of the oral area are available, and so extraoral facial movement generally becomes more important (and its influence more apparent) when oral area cues are unavailable through occlusion.

Access to subtle cues to visual speech outside the oral area also helps explain the different patterns of tolerance observed for visual and audiovisual speech perception when lower areas of the face were occluded. Visual speech influences on auditory speech perception (congruent and incongruent) remained substantial for lower-left and lower-right occluded displays but were reduced (although still evident for some speech stimuli) for lower-occluded displays. In contrast, visual speech perception ("speechreading") across lower-occluded, lower-left occluded, and lower-right occluded displays showed only an overall reduction in accuracy that did not vary across these displays or speech stimuli. These different patterns of tolerance of occlusion indicate different sensitivities of visual and audiovisual speech perception to visual speech outside the lower facial area and may reflect the way in which visual speech is processed perceptually. In particular, representa-

tions for visual speech across the face are likely to develop from natural exposure to audiovisual speech (since unimodal visual speech is normally rare; for a discussion, see Jordan & Thomas, 2001). Moreover, whereas explicitly identifying the auditory component of audiovisual speech is a normal and common occurrence, explicitly identifying visual speech is not (e.g., Aiello, 1991; Laver & Hutcheson, 1972) and is arduous even when it is an exceptionally important source of speech information (e.g., for deaf observers; Walden et al., 1977). So, whereas explicit access to visual speech representations may be effortful, implicit access for processing audiovisual speech may be more fluent (see also Massaro, 1987; Summerfield & McGrath, 1984) and may allow for a richer use of visual cues for audiovisual speech perception when they exist, including subtle cues from outside the oral area (see, e.g., Jordan & Thomas, 2001; Thomas & Jordan, 2004).

This would explain the tolerance of occlusion observed for audiovisual stimuli in lower-left occluded and lower-right occluded displays, where audiovisual speech perception could be influenced by subtle movement of the cheeks and upper jaw. However, even fluent access to audiovisual representations has its limits, and the availability of these subtle cues was much reduced in lower-occluded displays, although some influence on auditory perception still remained. In contrast, when speechreading, only an overall drop in explicit identification of unimodal visual speech was observed when the more prominent cues to visual speech were occluded in lower-occluded, lower-left occluded, and lower-right occluded displays. It seems that the subtle visual cues that were available for audiovisual speech perception in lower-right occluded and lower-left occluded displays and that showed selective effects for different speech stimuli in lower-occluded displays were not available for speechreading.

However, irrespective of the type of occlusion that was present when unimodal visual speech stimuli were shown, all speechreading errors across all display types across both experiments contained the same vowel as the stimulus, even when visual speech perception was most impaired by occlusion (i.e., when lower areas of the face were occluded). Thus, while consonant identities were sometimes lost to occlusion, vowel identities remained robust. The use of just two vowel identities (/a/ and /i/) may have contributed to this effect. Nevertheless, this finding suggests that while lower facial areas (when visible) provide cues that are sufficient for vowel identification, the different cues available in upper facial areas are able to provide the same levels of discrimination. Further work on vowel discrimination under conditions of occlusion will clarify the issue, but the indication is that cues supporting visual speech perception can be sufficiently spatially diverse across a talking face to overcome the reduced visibility produced by simple substantial occlusions.

The principal aim of the experiments reported in this article was to investigate the influences of simple substantial occlusion on visual and audiovisual speech perception. In common with previous studies of visual and audiovisual speech, the same talker was used in the construction of all displays. Unlike in most previous studies, however, this talker was selected to be representative of many talkers on the basis of facial movement, and so there is good reason to be confident about the generalizability of the findings. However, the tolerance of simple substantial occlusions that we observed may combine with other factors during normal social interactions. For example, tolerance of occlusion is likely to be augmented by other sources of information, particularly contextual, semantic, and syntactic, when they are available. In addition, the sustained presence of a particularly disruptive occlusion of a talker may shift the weighting an observer attaches to particular facial areas. For example, a beard may occlude much of the lower face, and the relevance of other facial areas (e.g., upper jaw/upper cheeks) may become greater than for other talkers. Moreover, speech may not be the only facial signal to tolerate substantial occlusion during social interaction, since the audibility of other socially significant facial signals, such as laughter, is also affected by seeing facial movement (Jordan & Abedipour, 2010). Indeed, laughter is often regarded as an evolutionary precursor to speech (e.g., Darwin, 1872; Niemitz, Loi, & Landerer, 2000; Provine, 2004), which may suggest that the tolerance of facial occlusion shown by speech perception has its origins at an evolutionarily earlier stage of human social communication that adapted to the demands of bimodal speech perception that eventually evolved.

In sum, the findings of this study show that visual and audiovisual speech perception has considerable tolerance of simple substantial occlusions and occurs even when left, right, upper, and lower areas of the face are entirely occluded. This suggests that visual and audiovisual speech perception overcomes substantial losses in facial visibility in everyday life and uses cues to visual speech from across the face that, when combined with experience-based processes, help provide a robust and adaptable system of visual and audiovisual speech perception that can accommodate a range of visual impoverishments.

References

Aiello, J. R. (1991). Human spatial behavior. In D. Stokols & I. Altman (Eds.), *Handbook of environmental psychology* (Vol. 1, pp. 389–504). Melbourne, FL: Krieger.

Annett, M. (1970). A classification of hand preference by association analysis. *British Journal of Psychology, 61,* 303–321.

Bennett, H., Delmonico, R., & Bond, C. (1987). Expressive and perceptual asymmetries of the resting face. *Neuropsychologia, 25,* 681–687.

Benoit, C., Guiard-Marigny, T., Le Goff, B., & Adjoudani, A. (1996). Which components of the face do humans and machines best speechread? In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by humans and machines: Models, systems and applications* (150th ed., pp. 315–328). Berlin: Springer. NATO ASI Series F: Computer and System Sciences.

Berger, K. W., Garner, M., & Sudman, J. (1971). The effect of degree of facial exposure and the vertical angle of vision on speech-reading performance. *Teacher of the Deaf, 69,* 322–326.

Biederman, I., & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 1162–1182.

Biederman, I., & Gerhardstein, P. C. (1995). Viewpoint-dependent mechanisms in visual object recognition: Reply to Tarr and Bülthoff (1995). *Journal of Experimental Psychology: Human Perception and Performance, 21,* 1506–1514.

Braddick, O. (1988). Contours revealed by concealment. *Nature, 333,* 803–804.

Butler, S., Gilchrist, I. D., Burt, D. M., Perrett, D. I., Jones, E., & Harvey, M. (2005). Are the perceptual biases found in chimeric face processing reflected in eye-movement patterns? *Neuropsychologia, 43,* 52–59.

Campbell, R. (1982). Asymmetries in moving faces. *British Journal of Psychology, 73,* 95–103.

Campbell, R. (1986). The lateralization of lip-read sounds: A first look. *Brain and Cognition, 5,* 1–21.

Carbary, T. J., Almierigi, J. B., & Harris, L. J. (1999). The left visual hemispace bias for the perception of composite faces: A test of the difficulty discrimination hypothesis. *Brain and Cognition, 40,* 71–75.

Cohen, M. M., Walker, R. L., & Massaro, D. W. (1996). Perception of visual speech. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by humans and machines: Models, systems and applications* (150th ed., pp. 315–328). Berlin: Springer. NATO ASI Series, Series F: Computer and System Sciences.

Darwin, C. (1872). *Expression of emotions in man and animals.* London: Murray.

Davis, C., & Kim, J. (2006). Audio-visual speech perception off the top of the head. *Cognition, 100,* B21–B31.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research, 12,* 423–425.

Gailey, L. (1987). Psychological parameters of lip reading skill. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip reading* (pp. 115–141). Hillsdale, NJ: Erlbaum.

Gilbert, C., & Bakan, P. (1973). Visual asymmetry in perception of faces. *Neuropsychologia, 11,* 355–362.

Graves, R. (1983). Mouth asymmetry, dichotic ear advantage and tachistoscopic visual field advantage as measures of language lateralization. *Neuropsychologia, 21,* 641–649.

Graves, R., Goodglass, H., & Landis, T. (1982). Mouth asymmetry during spontaneous speech. *Neuropsychologia, 20,* 371–381.

Green, K. P., Kuhl, P. K., & Meltzoff, A. N. (1988). Factors affecting the integration of auditory and visual information in speech: The effect of vowel environment. *Journal of the Acoustical Society of America, 84,* 155.

Greenberg, H. J., & Bode, D. L. (1968). Visual discrimination of consonants. *Journal of Speech and Hearing Research, 11,* 869–874.

International Organization for Standardization. (2006). New York: American National Standards Institute.

IJsseldijk, F. J. (1992). Speechreading performance under different conditions of video image, repetition, and speech rate. *Journal of Speech and Hearing Research, 35,* 466–471.

Jordan, T. R., & Abedipour, L. (2010). The importance of laughing in your face: Influences of visual laughter on auditory laughter perception. *Perception, 39,* 1283–1285.

Jordan, T. R., & Bevan, K. (1997). Seeing and hearing rotated faces: Influences of facial orientation on visual and audiovisual speech recognition. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 388–403.

Jordan, T. R., McCotter, M. V., & Thomas, S. M. (2000). Visual and audiovisual speech perception with color and gray scale facial images. *Perception & Psychophysics, 62,* 1394–1404.

Jordan, T. R., & Paterson, K. B. (2009). Re-evaluating split-fovea processing in word recognition: A critical assessment of recent research. *Neuropsychologia, 47,* 2341–2353.

Jordan, T. R., & Paterson, K. B. (2010). Where is the evidence for split fovea processing in word recognition? *Neuropsychologia, 48,* 2782–2783.

Jordan, T. R., & Sergeant, P. C. (1998). Effects of facial image size on visual and audiovisual speech recognition. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech* (pp. 155–176). London: Taylor & Francis.

Jordan, T. R., & Sergeant, P. C. (2000). Effects of distance on visual and audiovisual speech recognition. *Language and Speech, 43,* 107–124.

Jordan, T. R., & Thomas, S. M. (2001). Effects of horizontal viewing angle on visual and audiovisual speech recognition. *Journal of Experimental Psychology: Human Perception and Performance, 27,* 1386–1403.

Jordan, T. R., & Thomas, S. M. (2007). Hemiface contributions to hemispheric dominance in visual speech perception. *Neuropsychology, 21,* 721–731.

Larr, A. L. (1959). Speechreading through closed circuit television. *Volta Review, 61,* 19–21.

Laver, J., & Hutcheson, S. (1972). *Communication in face to face interaction: Selected readings.* Oxford: Oxford University Press.

Lindzey, G., Prince, B., & Wright, H. (1952). A study of human faces. *Journal of Personality, 21,* 687–694.

Lowe, D. G. (1987). The viewpoint consistency constraint. *International Journal of Computer Vision, 1,* 57–72.

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology, 12,* 131–141.

MacLeod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology, 24,* 29–43.

Marassa, L. K., & Lansing, C. R. (1995). Visual word recognition in two facial motion conditions: Full-face versus lips-plus-mandible. *Journal of Speech and Hearing Research, 38,* 1387–1394.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Erlbaum.

Massaro, D. W., & Cohen, M. M. (1996). Perceiving speech from inverted faces. *Perception & Psychophysics, 58,* 1047–1065.

McCurdy, H. G. (1949). Experimental notes on the asymmetry of the human face. *Journal of Abnormal Psychology, 44,* 553–555.

McGrath, M. (1985). *An examination of cues for visual and audiovisual speech perception using natural and computer generated faces.* Nottingham, UK: PhD Thesis, University of Nottingham.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.

Middleweerd, M. J., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *Journal of the Acoustical Society of America, 82,* 2145–2147.

Montgomery, A. A., & Jackson, P. L. (1983). Physical characteristics of the lips underlying vowel lipreading performance. *Journal of the Acoustical Society of America, 73,* 2134–2144.

Munhall, K. G., & Vatikiotis-Bateson, E. (1998). The moving face during speech communication. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech* (pp. 123–139). London: Psychology Press.

Nakayama, K., Shimojo, S., & Silverman, G. H. (1989). Stereoscopic depth: Its relation to image segmentation, grouping and the recognition of occluded objects. *Perception, 18,* 55–68.

Nicholls, M. E., Searle, D. A., & Bradshaw, J. L. (2004). Read my lips: Asymmetries in the visual expression and perception of speech revealed through the McGurk effect. *Psychological Science, 15,* 138–41.

Niemitz, C., Loi, M., & Landerer, S. (2000). Investigations on human laughter and its implications for the evolution of hominid visual communication. *Homo, 51,* 1–18.

Petajan, E. D. (Ed.). (1984). *Automatic lipreading to enhance speech recognition. In Proceedings of the IEEE Communications Society Global Telecommunications Conference, Atlanta, Georgia* (pp. 265–272). Piscataway, NJ: IEEE Press.

Preminger, J. E., Lin, H. B., Payen, M., & Levitt, H. (1998). Selective visual masking in speechreading. *Journal of Speech, Language, and Hearing Research, 41,* 564–575.

Provine, R. R. (2004). Laughing, tickling, and the evolution of speech and self. *Current Directions in Psychological Science, 13,* 215–218.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–113). Hillsdale, NJ: Erlbaum.

Ronneberg, J. (1993). Cognitive characteristics of skilled tactiling: The case of G. S. *European Journal of Cognitive Psychology, 5,* 19–33.

Rosenblum, L. D., Johnson, J., & Saldaña, H. M. (1996). Visual kinematic information for embellishing speech in noise. *Journal of Speech and Hearing Research, 39,* 1159–1170.

Rosenblum, L. D., Yakel, D. A., & Green, K. P. (2000). Face and mouth inversion effects on visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 806–819.

Samuelsson, S., & Ronneberg, J. (1993). Implicit and explicit use of scripted constraints in lip reading. *European Journal of Cognitive Psychology, 5,* 201–233.

Scheinberg, J. C. (1980). Analysis of speechreading cues using an interleaved technique. *Journal of Communication Disorders, 13,* 489–492.

Seinen, M., & Van der Werff, J. J. (1969). The perception of asymmetry in the face. *Nederlands Tijdschrift voor de Psychologie en Haar Grensgebieden, 24,* 551–558.

Sekuler, A. B. (1994). Local and global minima in visual completion—Effects of symmetry and orientation. *Perception, 23,* 529–545.

Sekuler, A. B., Palmer, S. E., & Flynn, C. (1994). Local and global processes in visual completion. *Psychological Science, 5,* 260–267.

Stone, L. (1957). *Facial clues of context in lip reading (John Tracy Clinic, Los Angeles, Research Papers 5).* Los Angeles: John Tracy Clinic.

Sudman, J. A., & Berger, K. W. (1971). Two-dimension vs. three dimension viewing in speech reading. *Journal of Communication Disorders, 4,* 195–198.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212–215.

Summerfield, A. Q. (1979). Use of visual information in phonetic perception. *Phonetica, 36,* 314–331.

Summerfield, A. Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3–51). Hillsdale, NJ: Erlbaum.

Summerfield, A. Q. (1992). Lipreading and audio-visual speech perception. In V. Bruce, A. Cowey, A. W. Ellis, & D. I. Perrett (Eds.), *Processing the facial image* (pp. 71–78). Oxford: Oxford University Press.

Summerfield, A. Q., MacLeod, P., McGrath, M., & Brooke, N. M. (1989). Lips, teeth and the benefits of lipreading. In A. W. Young & W. Ellis (Eds.), *Handbook of research on face processing* (pp. 223–233). Amsterdam: North-Holland.

Summerfield, A. Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology, 36A,* 51–74.

Thomas, S. M., & Jordan, T. R. (2002). Determining the influence of Gaussian blurring on inversion effects with talking faces. *Perception & Psychophysics, 64,* 932–944.

Thomas, S. M., & Jordan, T. R. (2004). Contributions of oral and extra-oral facial motion to visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 30,* 873–888.

Vatikiotis-Bateson, E., Munhall, K. G., Hirayama, M., Lee, Y. V., & Terzopoulos, D. (1996). Dynamics of facial motion in speech: Kinematic and electromyographic studies of orofacial structures. In D. G. Stork & M. Hennecke (Eds.), *Speechreading by humans and machines* (Vol. 150, pp. 221–232). Berlin: Springer. NATO-ASI Series F: Computer and System Sciences.

Vetter, T., Poggio, T., & Bülthoff, H. H. (1994). The importance of symmetry and virtual views in three-dimensional object recognition. *Current Biology, 4,* 8–23.

Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., & Jones, C. J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research, 20,* 130–145.

Wolff, W. (1933). The experimental study of forms of expression. *Character and Personality, 2,* 168–176.

Wolf, M. E., & Goodale, M. A. (1987). Oral asymmetries during verbal and non-verbal movements of the mouth. *Neuropsychologia, 25,* 375–396.

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics, 30,* 555–568.

Yehia, H. C., Rubin, P. E., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26,* 23–44.

Yokota, M. (1994). Perception of partly occluded objects in schizophrenic patients. *Japanese Psychological Research, 36,* 169–177.