Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect

MARTIN PARÉ, REBECCA C. RICHLER, and MARTIN ten HOVE
Queen's University, Kingston, Ontario, Canada

and

K. G. MUNHALL
Queen's University, Kingston, Ontario, Canada
and ATR Human Information Science Laboratories, Kyoto, Japan

We conducted three experiments in order to examine the influence of gaze behavior and fixation on audiovisual speech perception in a task that required subjects to report the speech sound they perceived during the presentation of congruent and incongruent (McGurk) audiovisual stimuli. Experiment 1 showed that the subjects' natural gaze behavior rarely involved gaze fixations beyond the oral and ocular regions of the talker's face and that these gaze fixations did not predict the likelihood of perceiving the McGurk effect. Experiments 2 and 3 showed that manipulation of the subjects' gaze fixations within the talker's face did not influence audiovisual speech perception substantially and that it was not until the gaze was displaced beyond 10º–20º from the talker's mouth that the McGurk effect was significantly lessened. Nevertheless, the effect persisted under such eccentric viewing conditions and became negligible only when the subject's gaze was directed 60º eccentrically. These findings demonstrate that the analysis of high spatial frequency information afforded by direct oral foveation is *not* necessary for the successful processing of visual speech information.

In its most natural form, human communication involves face-to-face conversation between two or more participants. Although acoustic speech is the primary medium of this communication, natural conversations include considerable visual information that is produced and processed in parallel with speech. Manual gestures, head and body posture, facial expression, changes in gaze direction, and visible speech movements convey an information-rich message that is combined with spoken language by the listener (McNeill, 1992). Identifying this visual information and understanding its role in communication is the focus of a broad interdisciplinary research effort (for a review, see Stork & Hennecke, 1996). The present study addresses one part of this large problem: the perception of visible speech movements. In particular, we examine the influence of gaze behavior and fixation on audiovisual speech processing.

It is well known that the sight of a talker's face can influence the perception of speech sounds: The intelligibility of speech in noise is enhanced when the listener can see the talker articulating (Sumby & Pollack, 1954). A similar type of audiovisual enhancement is an important part of rehabilitation for the hearing impaired, and some profoundly hearing-impaired individuals can seemingly substitute vision for auditory speech perception (Bernstein, Demorest, & Tucker, 2000); silent speechreading accuracy in some of these individuals can exceed 90%. Finally, visible speech can even influence the perception of perfectly audible speech sounds. In their classic study, McGurk and MacDonald (1976) showed that auditory perception is modified when a face articulating one sound is dubbed in synchrony with a voice producing another sound. This McGurk effect is typically produced by the combination of auditory and visual sources that differ in place of articulation—for example, a visual (velar) /g/ dubbed onto an acoustic (bilabial) /b/ frequently yields the perception of a different consonant, /d/. This audiovisual illusion has been replicated numerous times, and some of the conditions influencing its perception have been identified (e.g., Green & Kuhl, 1989; Green, Kuhl, Meltzoff, & Stevens, 1991; MacDonald & McGurk, 1978; Massaro, 1987; Massaro & Cohen, 1993; Munhall, Gribble, Sacco, & Ward, 1996; Sekiyama & Tohkura, 1991; Summerfield & McGrath, 1984). In the present study, we used the

McGurk effect as a tool to study the influence of gaze behavior on audiovisual speech processing.

Monitoring eye movements has been an important technique in the study of a number of perceptual and cognitive tasks, including picture and object perception (Yarbus, 1967), reading (Rayner, 1998), chess playing (Reingold, Charness, Pomplun, & Stampe, 2001), and spoken word recognition (Tanenhaus, Magnuson, Dahan, & Chambers, 2000). The rationale behind this approach is that the patterns of gaze fixations reveal active information-gathering routines that reflect "expert" knowledge of the stimuli and tasks.

To date, very few studies have investigated gaze behavior in audiovisual speech perception. In one study that used a speech-in-noise paradigm, Vatikiotis-Bateson and colleagues (Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998) found that gaze fixations alternated between the eyes and the mouth, with no change in this pattern as a function of image size and with only modest changes as the auditory signal-to-noise level varied. As the signal-to-noise level decreased, subjects tended to fixate the mouth more often, but even at the highest noise level the percentage of gaze fixations at the mouth represented only 55% of all fixations. In another study, Lansing and McConkie (1999) examined gaze behavior in silent speech-reading and observed differences in gaze behavior when the speech recognition task emphasized recognition of individual words versus determining the intonation of the sentence. Subjects preferentially fixated lower facial regions (the mouth) when identifying individual words and upper facial regions (the eyes) when carrying out the intonation tasks. Although these studies demonstrated systematic patterns of gaze behavior during audiovisual speech perception, it remains unclear to what extent gaze fixations on a talker's face influence perception. On the one hand, the motion of articulation spreads across the face (Vatikiotis-Bateson, Munhall, Hirayama, Kasahara, & Yehia, 1996), and these distributed motions correlate strongly with the acoustics of the speech (Munhall & Vatikiotis-Bateson, 1998; Yehia, Rubin, & Vatikiotis-Bateson, 1998), indicating that foveating facial features distant from the mouth still allows redundant visual information to be gathered. On the other hand, it has also been demonstrated that visual speech processing is possible even when high spatial frequency information is eliminated from the images (e.g., C. S. Campbell & Massaro, 1997; MacDonald, Andersen, & Bachmann, 2000), indicating that low spatial frequency analysis performed by parafoveal and peripheral vision could be sufficient for the processing of visible speech information. Thus, it could be that direct foveation is not required for successful visual speech processing.

In this research, we conducted three experiments in order to examine the influence of gaze behavior and fixation on audiovisual speech perception in a task that required subjects to report the speech sound they perceived during the presentation of congruent and incongruent (McGurk) audiovisual stimuli. In the first experiment, we examined the unconstrained gaze behavior of subjects performing this audiovisual speech perception task and tested whether different gaze fixations influenced the occurrence of the McGurk effect. In the second experiment, we systematically studied the influence of gaze fixations on audiovisual speech perception by instructing the subjects to direct their gaze to specific facial regions of the talker uttering the same congruent and incongruent audiovisual stimuli. In the third experiment, the influence of gaze fixations on audiovisual speech perception was further investigated, with the subjects instructed to fixate on spots of lights beyond the limits of the talker's face.

GENERAL METHOD

Subjects

A total of 61 subjects participated in these experiments. Most of the subjects were undergraduate students in the Department of Psychology at Queen's University, and they were naive about the experimental hypotheses. All the subjects were native speakers of Canadian English and were required to have normal or corrected-to-normal (20/20) vision, as well as no hearing or speech difficulties (established through self-report). They were paid $10 for participating. All experimental protocols were approved by Queen's University Research Ethics Board.

Stimuli

The stimuli for the experiments were professionally filmed using digital audio and video recording equipment. They consisted of a series of five nonsense utterances—/aba/, /ada/, /aga/, /aða/, and /ava/—produced by four fluent speakers of English: three females (Talkers 1, 2, and 4) and one male (Talker 3). Each talker was filmed against a blue background, framed so that only the head and shoulders were visible (see Figure 1). The videotape was then pressed onto a laser disc to permit randomization of the talkers and stimuli during testing. This format allowed for repeated use without degradation in audiovisual quality.

Figure 1. Gaze calibration trials of one of the face images used in the experiments. The two-dimensional position of the subject's dominant eye sequentially fixating on the mouth and the right and left eyes of the static face stimulus is shown. The *windows* (4° × 4°) centered on the average gaze fixation positions indicate the mouth and eye regions used in the data analysis.

The acoustics of the five utterances for each talker were digitized with a PC using a Sound Blaster 16-bit A/D converter. The stimuli were analog-filtered at 20 kHz and were sampled at 44 kHz. These signals were subsequently used for the auditory-only test, as well as for the congruent and incongruent audiovisual tests. For the audiovisual condition, these utterances were synchronized with the visual stimuli by aligning the acoustic burst of the auditory stimulus with the acoustic burst in the soundtrack of the video disc. For each talker, two sets of stimuli were created: (1) congruent stimuli consisting of matching auditory and visual stimuli and (2) incongruent McGurk stimuli obtained by dubbing an auditory /aba/ onto each of the other visual stimuli. The timing of both the congruent and the incongruent audiovisual stimuli was controlled using custom software and was verified using an A/D system and a digital waveform editor. The audiovisual synchrony was reliably reproduced with 1-msec accuracy.

Apparatus

Unless specified otherwise, the subjects sat at a desk with their heads restrained in a head- and chinrest so that their eyes were posi-

tioned 114 cm from a 20-in. television monitor (Sony Model PVM 1910). The video disc was played with a Pioneer LD-8000 video disc player, with acoustics played through an MG Electronics Cabaret speaker, placed directly below the monitor. Custom software was used to control the video disc trials, synchronize the auditory stimuli with the video, and record the subjects' responses from a labeled keyboard. Unless specified otherwise, the experiments took place in a double-walled soundproof booth (IAC Model 1204).

Procedure

The subjects were brought into the experimental room and were seated in front of the labeled keyboard. Along with specific instructions particular to each experiment, they were instructed to "watch" the talker on the monitor at all times. A blank screen was presented following each stimulus presentation, and a response was required to cue the next utterance. We used block randomization to present the stimuli. The subjects' responses were recorded on the computer by use of a keyboard, which was labeled so that each of six keys in the middle row represented a possible response (b, th, v, d, or g), in-
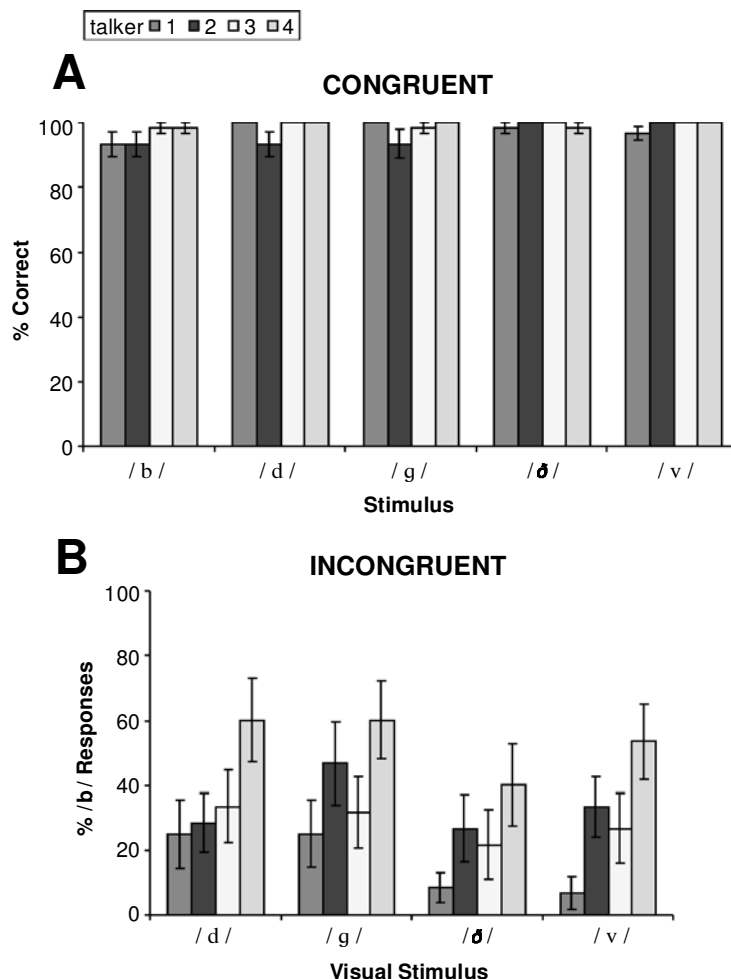

Figure 2. Accuracy performance in response to the presentation of the congruent (A) and incongruent (B) audiovisual stimuli. (A) Percentages (± *SE*) of correct responses. (B) Percentage of correct /b/ responses; a low percentage of /b/ responses indicates a strong McGurk effect. The results are from 12 participants, 2 males and 10 females, ranging from 18 to 23 years of age. A total of 180 trials was obtained: five consecutively played blocks, with each of the four talkers producing nine stimuli, of which five were congruent (/b/, /d/, /g/, /ð/, and /v/) and four incongruent (auditory /b/ dubbed onto visual /d/, /g/, /ð/, or /v/).

	Stimuli					
Modality	/b/	/d/	/g/	/ð/	/v/	Average
Audio	98.6	100.0	99.3	81.3	93.8	94.6
Visual	98.6	56.9	84.7	97.9	97.9	87.2

Note—Results are from 12 participants, 3 males and 9 females, from 18 to 19 years of age. A total of 120 trials was obtained for the auditory-only and visual-only conditions: three blocks with each of the four talkers producing five stimuli (/b/, /d/, /g/, /ð/, and /v/). Three audio-only and three visual-only blocks were played consecutively, with the order counterbalanced between subjects. The subjects were instructed to report what they heard during the audio-only trials and what the talker seemed to say during the visual-only trials.

cluding a key labeled *o* for anything other than the five expected responses. Key order was counterbalanced across subjects, and the order of stimulus presentation was randomized within each block of trials for each subject. The behavioral responses (what the subjects reported having heard) were categorized with respect to the auditory stimuli presented and were expressed in terms of percentage of correct responses. For trials in which incongruent stimuli were presented, the results were reported as the percentage of the trials that the auditory stimulus was correctly identified—that is, the percentage of correct /b/ responses.[1] Unless specified otherwise, data from the incongruent stimulus condition of each experiment were analyzed with an analysis of variance (ANOVA), followed by an all-pairwise multiple comparison procedure (Student–Newman–Keuls test). Such ANOVAs were not conducted on the response data from the congruent stimulus condition, because the high percentages of correct responses in this condition produced a ceiling effect that eliminated most of the data variance and precluded inferential statistics.

Stimulus Characteristics

We performed two preliminary experiments to assess the characteristics of the audiovisual stimuli. First, we tested the audio and the visual stimuli separately by asking 12 subjects to identify the tokens in two single-modality (auditory-only and visual-only) tasks. Table 1 shows the mean percentage of correct responses for each stimulus in the two modalities. These single-modality stimuli elicited familiar response patterns (e.g., Miller & Nicely, 1955). For example, the strong acoustic stop-consonants /d/ and /g/ yielded almost perfect accuracy in auditory-only trials but were easily confused when presented visually. In contrast, the subjects performed less accurately with the consonants /v/ and /ð/ presented solely with auditory information, but they were easily identified in visual-only trials. These results also revealed that the four talkers did not elicit equivalent responses, their different speech styles possibly creating diversity in the amount of phonetic information (Demorest & Bernstein, 1992; Gagné, Masterson, Munhall, Bilida, & Querengesser, 1995). In a second step, and with a new set of 12 subjects, we assessed the audiovisual characteristics of the stimuli and determined the respective effectiveness of each talker at producing the McGurk effect. Figure 2 shows the accuracy of the auditory performance for each of the congruent and incongruent stimuli and for each of the talkers. We found that the incongruent audiovisual stimuli were equally successful at eliciting the McGurk effect. The talkers uttering them were, however, not equally efficient at eliciting the effect; Talker 4's stimuli produced a significantly weaker effect.

EXPERIMENT 1

In the first experiment, we studied the natural visual information-gathering processes during audiovisual speech perception. The gaze behavior of subjects was monitored while they performed a consonant perception task involving the presentation of congruent and incongruent (McGurk) audiovisual stimuli. We examined the significance of foveation during audiovisual speech perception by relating the perceptual responses and the gaze fixation positions during each individual trial. If oral foveation is important for the successful integration of audio and visual speech information, it is hypothesized that the strength of the McGurk effect will vary as a function of how close the subject's gaze is to the talker's mouth.

Method

Subjects. Nine new subjects, 4 male and 5 females, ranging from 22 to 35 years of age, were tested.

Stimuli. One hundred eighty trials were obtained: five consecutive blocks, with each of the four talkers producing nine stimuli, of which five were congruent (/b/, /d/, /g/, /ð/, and /v/) and four were incongruent (auditory /b/ dubbed onto visual /d/, /g/, /ð/, or /v/).

Apparatus. This experiment was performed in a quiet, but not soundproof, laboratory room. The visual stimuli were presented on a 37-in. video monitor (Mitsubishi XC-3730C Megaview Pro 37), and the acoustics were played through the monitor speakers, which were positioned on each side of the monitor. We used the search-coil-in-magnetic-field technique (Robinson, 1963) to monitor the gaze position of each of the 9 subjects by positioning them in the center of a 31-in. field coil.

Procedure. During the experiment, the subjects were seated in a chair with their heads restrained in a head- and chinrest so that their eyes were situated in the middle of the field coil. They faced the video monitor, which was positioned 114 cm in front of their eyes. Each subject was instructed to listen to each stimulus while watching the talker on the video monitor and to respond strictly with what he or she heard. The induction coil consisted of a light coil of wire that was embedded in a flexible ring of silicone rubber (Skalar) that adhered to the limbus of the human eye, concentric with the cornea (Collewijn, van der Mark, & Jansen, 1975). Before this search coil was put into place, an ophthalmologist examined each subject to ensure that he or she did not show signs of ocular pathology causing a peripheral visual field defect and that his or her central visual acuity was not less than 20/20. The search coil was positioned in the dominant eye of the subjects only after the surface of the eye had been anesthetized with a few drops of anesthetic (Tetracaine HCl, 0.5%).

Prior to the experimental block, calibration of the search coil was obtained by instructing the subjects to fixate on discrete spots of light regularly distributed across the visual working space. Further calibration trials were run following the experiment. These trials consisted of the presentation of a static image of each of the four talkers, during which the subjects were instructed to fixate alternatively on the mouths and each of the eyes of the talkers (see Figure 1).

Data acquisition and analysis. Along with the acoustic signal from the video disc player, the demodulated voltages of the search coil encoding the horizontal and vertical components of the eye position were filtered (low-pass, anti-aliasing, 8-pole Bessel filter with a frequency corner of 241 Hz), digitized to a resolution of 12 bits, and sampled at 1 kHz (A/D converter; Analogic) by a Pentium III computer running a UNIX-based real-time experimentation software (Hays, Richmond, & Optican, 1982). Following the completion of the experimental sessions, these data were displayed trial by trial on a computer running an interactive software program that allowed the marking of the nearest gaze fixation positions at the time of occurrence of the first (~500 msec of stimulus presentation) and second (~1 sec of stimulus presentation) vowels. The gaze data were then combined with the behavioral responses off line, sorted according to the audiovisual stimulus congruence, the talker, the stimulus, and the perceptual response. To analyze the spatial distributions of the gaze fixations, we used the calibration data to divide the

talker ■1 ■2 □3 □4

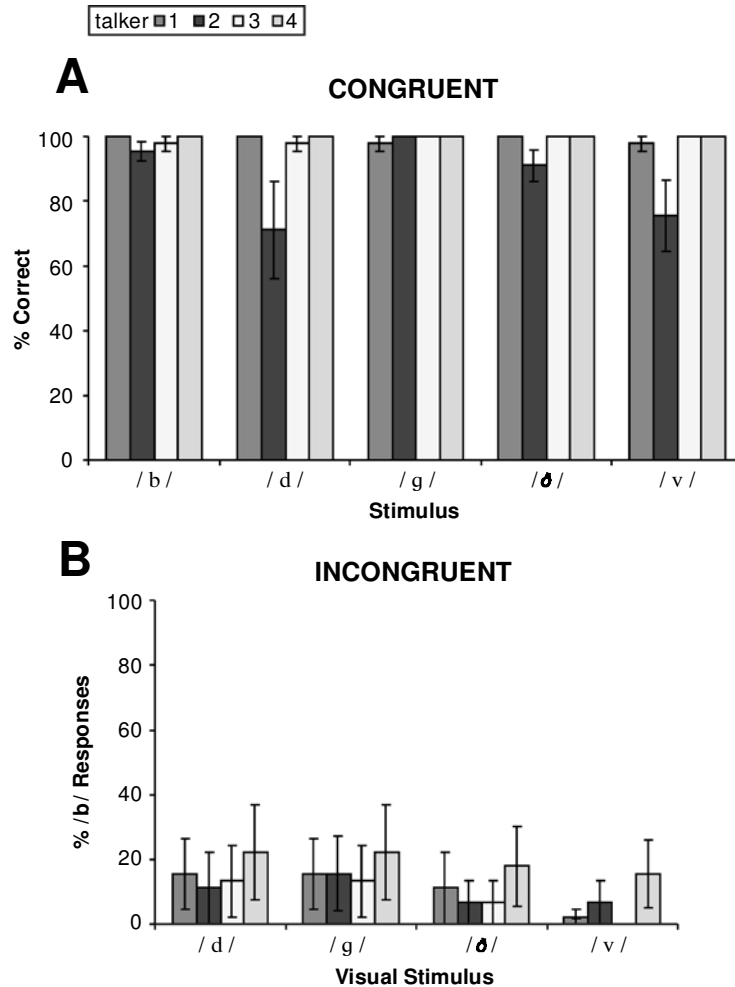**A**   **CONGRUENT**

B   **INCONGRUENT**

Figure 3. Accuracy performance in response to the presentation of the congruent (A) and incongruent (B) audiovisual stimuli produced by the four talkers for the 9 subjects in Experiment 1.

talker's face into three regions: the mouth and each of the eyes. The subjects were assumed to be fixating one of these identified regions if the gaze fixation position was within a 4° × 4° *window* centered on the calibrated horizontal and vertical positions (Figure 1). These regions captured all the gaze fixations clearly directed to the eyes and the mouths of the talkers.

Results and Discussion

Figure 3 shows, for each of the talkers, the mean percentages of correct responses for each of the congruent (A) and incongruent (B) stimuli. Data for the incongruent stimulus condition were analyzed with a two-way ANOVA. Neither the main effect of stimulus [$F(3,24) = 0.9$, $p = .44$] nor that of talker [$F(3,24) = 0.93$, $p = .43$] was statistically significant; note that Talker 4 still elicited a weaker McGurk effect. The average ($\pm SE$) percentage of /b/ responses for the incongruent stimuli was 12.2% ± 2.5%,[2] whereas the average ($\pm SE$) percentage of correct responses for the congruent stimuli was 96.2% ± 1.1%.

In both the congruent and the incongruent audiovisual stimulus conditions, the gaze behavior of each subject was found to be fairly stereotypical, with the gaze fixations being limited to very few facial features of the talkers. Figure 4 exemplifies this behavior by illustrating the gaze position traces, as a function of time, of 1 subject viewing the same talker in the incongruent (top) and the congruent (bottom) conditions. Following the onset of the stimuli presentation, a single saccade usually brought the subject's gaze onto a facial feature, such as one of the eyes or the mouth, and only a few fixation corrections were thereafter made. We analyzed the gaze behavior of each subject by measuring the gaze fixation positions at the acoustic onset of the first and second vowels, the perception of the second vowel being essential for the occurrence of the McGurk effect. Figure 4 additionally shows the distribution of these vowel-associated gaze fixations for the subject whose gaze position traces are illustrated. It appears

INCONGRUENT STIMULI
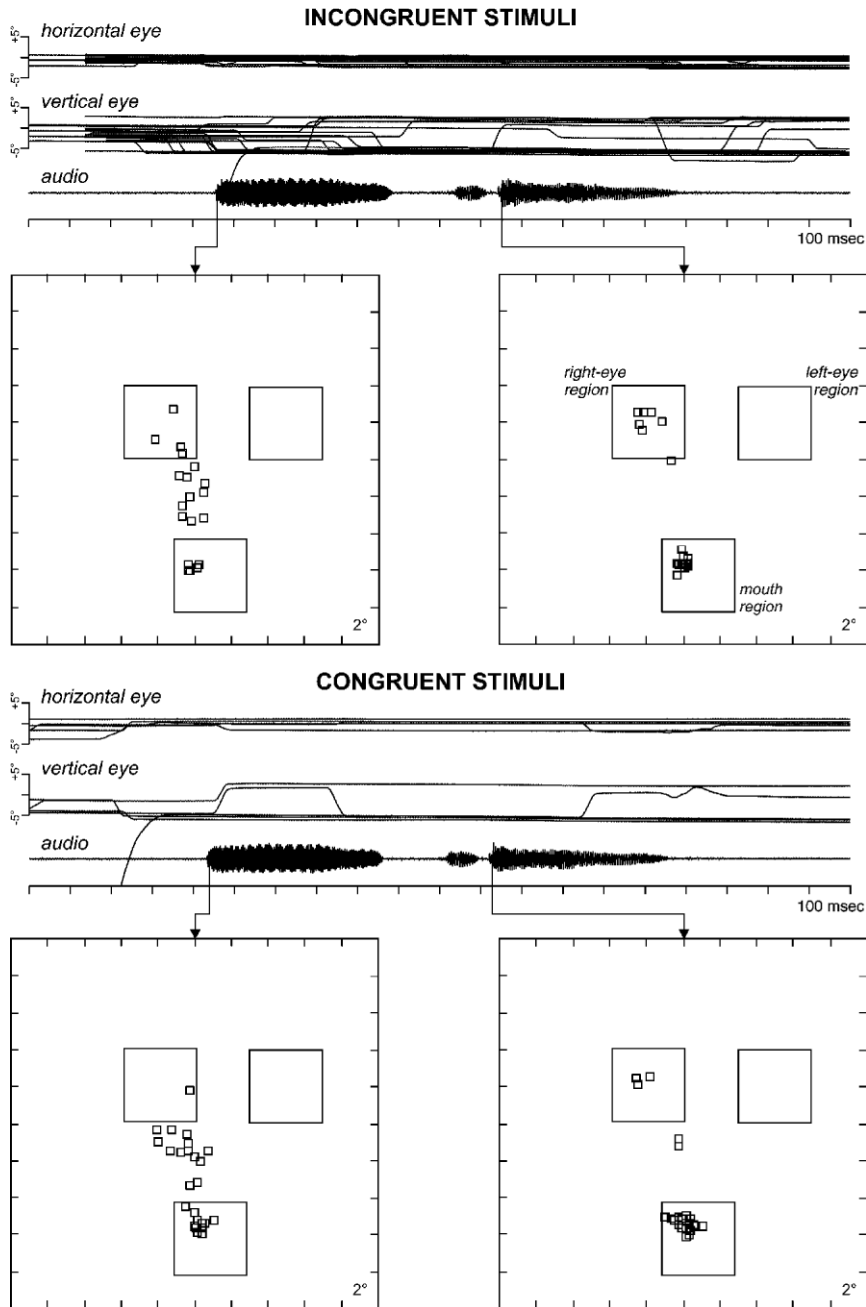

CONGRUENT STIMULI

Figure 4. Gaze behavior during the presentation of the incongruent (top) and congruent (bottom) audiovisual stimuli. Each panel shows the horizontal and vertical eye position traces (along with the audio signal /aba/), as a function of time, for 1 subject viewing one talker (4). The horizontal (abscissa) and vertical (ordinate) positions of the gaze fixations at the time of the onset of the first and second vowels (indicated by the arrow lines) are additionally shown in the corresponding graphs. Data from all the congruent and incongruent trials are presented, except for the eye position traces of the congruent trials, which show only those that corresponded with the acoustic /b/ stimulus.

that, from the first to the second vowel, the subject's gaze narrowed itself into clusters centered on either the mouth or the right eye of the talker. A very similar *narrowing* was observed in all the subjects viewing any of the talkers, as well as in both the incongruent and the congruent stimulus conditions. Even in the few cases in which the gaze fixations at the onset of the first vowel were concentrated on a single facial feature of the talker (e.g., the mouth), the variance of the gaze fixation positions was still reduced at the onset of the second vowel (not shown).

The vowel-associated gaze fixation positions were quantitatively analyzed by determining the proportions of trials on which the subjects fixated on the mouth, on either of the two eyes, or on any other positions. Figure 5 summarizes the results of this analysis for the congruent and the incongruent data. At the time of the first vowel's onset, many gaze fixations were on the mouths (44.7%), and few were on the talkers' eyes; the latter were predominantly more on the right eyes (14.7%) than on the left (1.8%). The rest of the gaze fixations were at some other positions (38.8%), with nearly all of them centered on the faces of the talkers—that is, around and on their noses. At the time of the onset of the second vowel, the majority, but not all, of the gaze fixations were then on the mouths (62.1%), and when the gaze was directed at the eyes (11.0%) the fixations were again concentrated on the right eyes of the talkers (9.8%). Nearly all the other gaze fixations (26.9%) were at intermediate central positions. The differences between the gaze fixation proportions at the onsets of the first and the second vowels were statistically significant whether we considered the congruent [$\chi^2(3) = 58.2, p < .001$] or the incongruent [$\chi^2(3) = 43.4, p < .001$] condition. Thus, the subjects fixated on the talkers' mouths more frequently, albeit not exclusively, at the time of the second vowel than at earlier times in the trial. Between congruent and incongruent conditions, a small significant difference between the proportions of gaze fixation positions at the second vowel was detected [$\chi^2(3) = 8.9, p < .05$], but not for the proportions at the first vowel [$\chi^2(3) = 2.9, p = .40$]. This effect was caused by an increase of right eye fixations at the expense of mouth fixations. In summary, the subjects adjusted their gazes over the course of individual trials and fixated predominately on the mouths of the talkers at the time of the second audiovisual

vowel, whether congruent or incongruent. When not fixating on the mouths, they tended to fixate either on the right eyes of the talkers or on some locations close to their noses.

Does the gaze behavior in incongruent conditions differ when the McGurk effect is perceived from when it is not? To answer this question, we again analyzed the distribution of gaze fixation positions. Because of the generally small number of non-McGurk trials (see above), the data from individual subjects could only be examined qualitatively. Figure 6 shows the gaze fixation positions, at the time of the second vowel, of 1 subject viewing two different talkers in the incongruent condition when the McGurk effect was (■) or was not (□) perceived. Obvious differences could not be detected from these individual data sets. Of the 9 subjects, 4 experienced the McGurk effect in every incongruent stimulus trial. Table 2 gives the different proportions of gaze fixation positions of the 5 subjects whose responses to the incongruent stimuli varied across trials. Although the proportion of fixations on the mouth was greater when the effect was perceived, there was no significant benefit from fixating the mouth. A chi square testing whether the proportion of mouth fixations (relative to all the other fixations) was larger when the effect was perceived than when it was not revealed no statistical difference [$\chi^2(1) = 0.8, p = .37$]. The proportion of mouth fixations (relative to all the other fixations) for the correct perception of the congruent stimuli (544/861, or 63.2%) also was not larger than that for the incorrect trials [26/39, or 66.7%; $\chi^2(1) = 0.07, p = .79$].

These data indicate that the gaze behavior of the subjects presented with the image of a talker uttering either congruent or incongruent audiovisual stimuli was rather stereotyped and that it did not correlate with the percep-
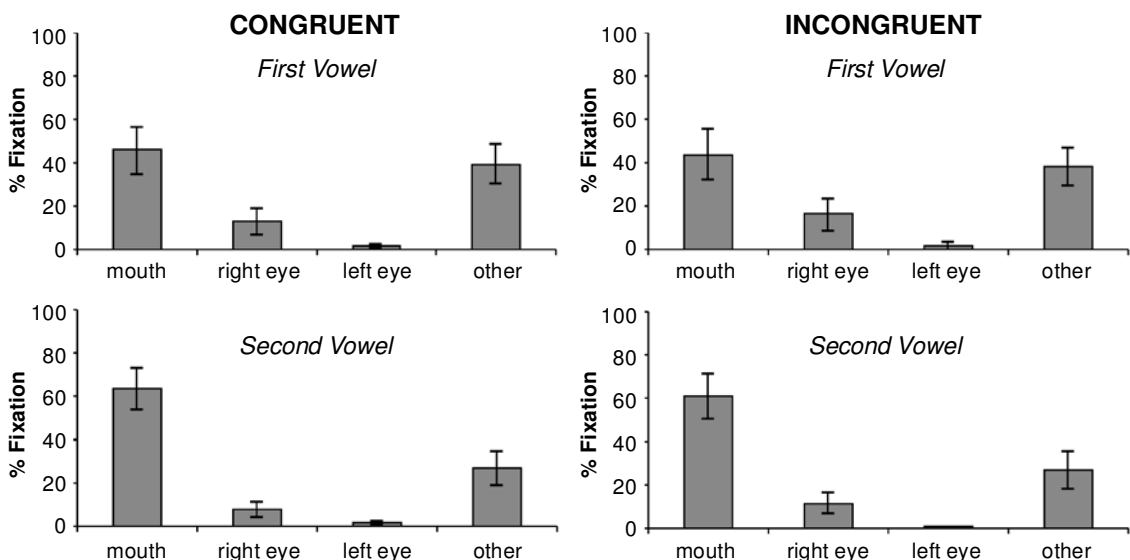

Figure 5. Percentages of gaze fixations on salient facial features of the talkers at the onset of the first and second vowels of the audiovisual stimuli for both the congruent and the incongruent stimuli.
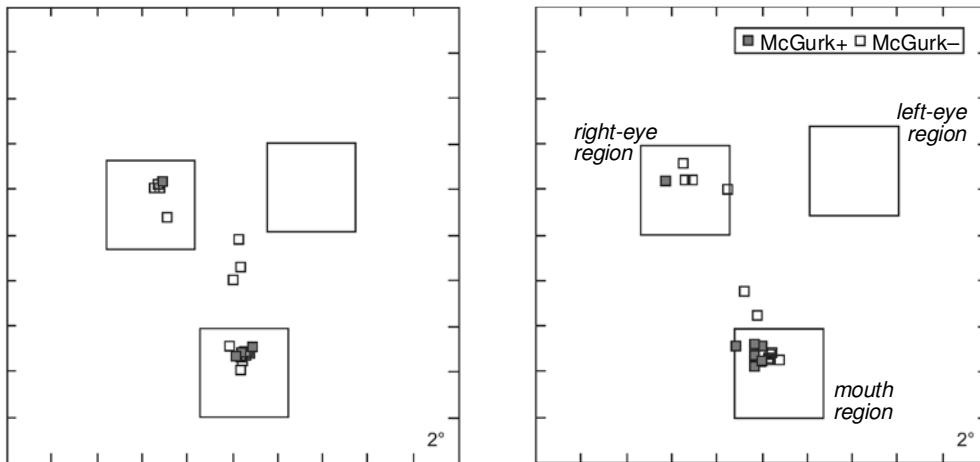
Figure 6. Horizontal (abscissa) and vertical (ordinate) gaze fixation positions of a single subject at the time of the onset of the second vowel of all the incongruent stimuli uttered by a single talker. Trials are identified according to whether the subject did (solid symbols) or did not (open symbols) perceive the McGurk effect. The left and right graphs display trials associated with Talkers 2 and 3, respectively.

tion of the McGurk effect. The subjects chose to fixate predominantly, but not exclusively, on the mouth of the talker during this simple audiovisual speech perception task. The perception of the McGurk effect, however, was not significantly enhanced by these fixations. Apart from the mouth, the other privileged gaze fixations in this experiment were either on the right eye of the talker or somewhere close to the nose. Indeed, a large proportion (39%) of the initial gaze fixations were, surprisingly, directed to the central facial region of the talker, and fixations in this region still represented 27% of the fixations at the time of the second vowel onset. Nevertheless, such fixations did not influence the perception of the McGurk effect. These results suggest that oral fixations are not necessary for the integration of audiovisual speech information to occur. One possibility is that the parafoveal/peripheral visual processes are sufficient for capturing *global* visual speech information. Alternatively, extra-oral fixations may adequately collect *local* visual speech information from the correlated motion usually distributed across the talker's face (Munhall & Vatikiotis-Bateson, 1998; Yehia et al., 1998). In this respect, the central (nose) fixations are interesting, since they could represent an ideal vantage point allowing foveal and parafoveal vision to process speech information from the talker's mouth and face concurrently.

Although the main targets of fixation (eyes, nose, and mouth) correspond to those reported previously (Lansing & McConkie, 1999; Vatikiotis-Bateson et al., 1998), the distribution of time spent fixating on the different facial locations differs from the distributions in those studies. Only 11% of all the fixations in the present study were spent on the eyes. Vatikiotis-Bateson et al. (1998) reported that 45%–70% of all fixations during the perception of their extended monologues were on the region of the eyes. Lansing and McConkie found that 39.5% of the gaze fix-

ations during a prosodic discrimination task were within an upper facial (eyes) region, whereas these represented only 14.4% during silent speechreading. The last authors reported that the fixations were predominantly on the mouth during a speechreading task. Although there is no single reason for this variability in fixation behavior, a few explanations are obvious. One contribution to these large differences in distribution of gaze may reside in the operational definition of the eye and mouth regions. At one end, the eye region in the study of Vatikiotis-Bateson et al. (1998) included part of the nasal area of the talker's face. At the other end, the eye region in this study closely matched the dimension of the talker's eye (see Figure 1). Lansing and McConkie's eye region was intermediate in size but included a region above the eyes. Nevertheless, such differences in region definition cannot account for all of the differences. For example, even if we combined the data for our *eye* and *nose* fixations, the combined data would account for only 37.9% of all the fixations, which is less

Table 2
Percentages of Gaze Fixations on Salient Facial Features
of the Talkers at the Onset of the Second Vowel
of the Audiovisual Stimulus

Condition	Facial Feature			
	Mouth	Right Eye	Left Eye	Other
Overall				
(*n* = 900)	75.3	6.9	0.1	17.7
Congruent				
(*n* = 500)	75.2	5.6	0.2	19.0
Incongruent				
(*n* = 400)	75.5	8.5	0.0	16.0
Incongruent, non-McGurk effect				
(*n* = 88)	70.5	15.9	0.0	12.5
Incongruent, McGurk effect				
(*n* = 312)	76.9	6.4	0.0	17.0

Note—Results are from the 5 subjects who experienced the McGurk effect.

than the eye fixations reported by Vatikiotis-Bateson et al. (1998).

A second and perhaps more important cause of these different distributions of fixations may reside in the different demands of the various experimental tasks. Listening to extended monologues in noise, silent speechreading, silent prosodic category discrimination, and audiovisual phoneme identification in nonsensical syllables requires vastly different levels of linguistic and social analyses. Although none of these laboratory tasks approaches the complexity of natural communication, some gaze strategies can be identified. Eye fixations are fairly typical in face recognition tasks (Yarbus, 1967), and recognition of speech intonation seems to promote them (Lansing & McConkie, 1999). The longer monologues in Vatikiotis-Bateson et al. (1998) also may have induced more eye fixations for social information. Monitoring the direction of another person's gaze is an important part of social interaction (Langton, Watt, & Bruce, 2000), and it is reasonable to assume that making social judgments about a person can influence the proportion of gaze fixations on the eyes. Similarly, time spent fixating the mouth has been shown to be slightly more frequent in poorer, more difficult listening conditions (Vatikiotis-Bateson et al., 1998). This adaptability in gaze behavior may suggest that subjects adopt the strategy of fixating the talker's mouth (or eyes) to maximize perception of particular information (Lansing & McConkie, 1999), but whether these gaze changes actually do or do not help specific visual speech information gathering cannot presently be determined. Our data seem to suggest the negative.

An additional question of interest is why were the ocular fixations of our subjects almost exclusively concentrated on the right eye of the talkers? Since language processing is localized primarily in the left hemisphere, one would have expected the subjects' gaze fixations to be biased toward their right visual field—that is, the left side of the talker's face. The behavioral strategy adopted by our subjects could, however, be explained by perceptual asymmetries in facial expressions biased toward the right side of a talking face (Burt & Perrett, 1997) or by an asymmetry in facial motion favoring the right upper hemiface (Richardson, Bowers, Bauer, Heilman, & Leonard, 2000). It may also be relevant that face processing is predominantly performed by the right cerebral hemisphere (Benton, 1990; Moscovitch, Scullion, & Christie, 1976; Rhodes, 1985).

The McGurk effect in this experiment was stronger than that in the second preliminary experiment. There are several possible explanations for this difference. First, the larger monitor that we used in the present experiment could have strengthened the visual stimuli, even though varying image size within the range of our experiments has been shown to have no effect on audiovisual speech perception (Jordan & Sergeant, 2000) and the associated gaze behavior (Vatikiotis-Bateson et al., 1998). Second, unlike the preliminary experiment, this one was not conducted in a soundproof laboratory, which most probably provided acoustic conditions of inferior quality; auditory

stimuli of poorer quality than the visual stimuli more frequently yield McGurk integration (Massaro, 1987). A third possibility is that the monitoring of the subject's gaze may have instigated a higher compliance from the subjects in fixating the monitor during every trial. Indeed, failure to watch the talker can potentially reduce the integration of the audiovisual stimuli, thereby yielding more purely auditory responses and a weaker McGurk effect. Finally, the intersubject variability of the McGurk effect and the relatively small samples of subjects studied could be an important factor. Although 4 out of 9 (44%) subjects perceived the effect on every trial in this experiment, there was only 1 such subject (1/12, or 8%) in the preliminary experiment.

EXPERIMENT 2

The results of Experiment 1 suggest that gaze fixations within the talker's face naturally adopted by the subjects did not influence the occurrence of the McGurk effect. This assertion, however, relied on a relatively small number of trials, wherein the subjects fixated predominantly the talkers' mouths and secondarily their eyes or noses. The present experiment systematically manipulated these natural gaze fixation tendencies in order to examine further whether gaze fixations on specific regions of the talker's face influence audiovisual speech perception. The subjects were thus specifically instructed to fixate on the mouth, on the eyes, or on the talkers' hairline during the presentation of congruent and incongruent (McGurk) audiovisual stimuli.

Methods

Subjects. Twelve female subjects between the ages of 18 and 20 years were tested.

Stimuli. A total of 324 trials were obtained—that is, 108 trials for each of the 3 gaze fixation positions consisting of three consecutive blocks with each of the four talkers, each producing nine stimuli, of which five were congruent (/b/, /d/, /g/, /ð/, and /v/) and four incongruent (auditory /b/ dubbed onto visual /d/, /g/, /ð/, or /v/).

Procedure. The subjects were seated 114 cm from the video monitor, with their head restrained in a head- and chinrest. They were instructed to maintain fixation[3] on one of the three locations on the talker's face: the mouth (mouth fixation condition), between the eyes (eye fixation condition), and the hairline (hairline fixation condition). The latter gaze fixation position corresponded to the facial position the farthest from the talker's mouth at the edge of the monitor. Its spatial location was marked with a small (0.1° diameter) tag. Across the talkers, the eyes were, on average, ~5° above the mouth, and the hairline was ~10° above the mouth. For each gaze fixation location, three blocks of trials were played consecutively. Following the completion of these three blocks, the subject was given a new gaze fixation to complete another three blocks. The order of fixations was counterbalanced across subjects. The subjects were instructed to report what they heard in each trial.

Results and Discussion

Figure 7 shows, for each of the three gaze fixation positions, the mean percentage of correct responses for each of the congruent (A) and incongruent (B) stimuli. The data for the incongruent stimulus condition were analyzed with a three-way ANOVA with stimulus, talker, and gaze
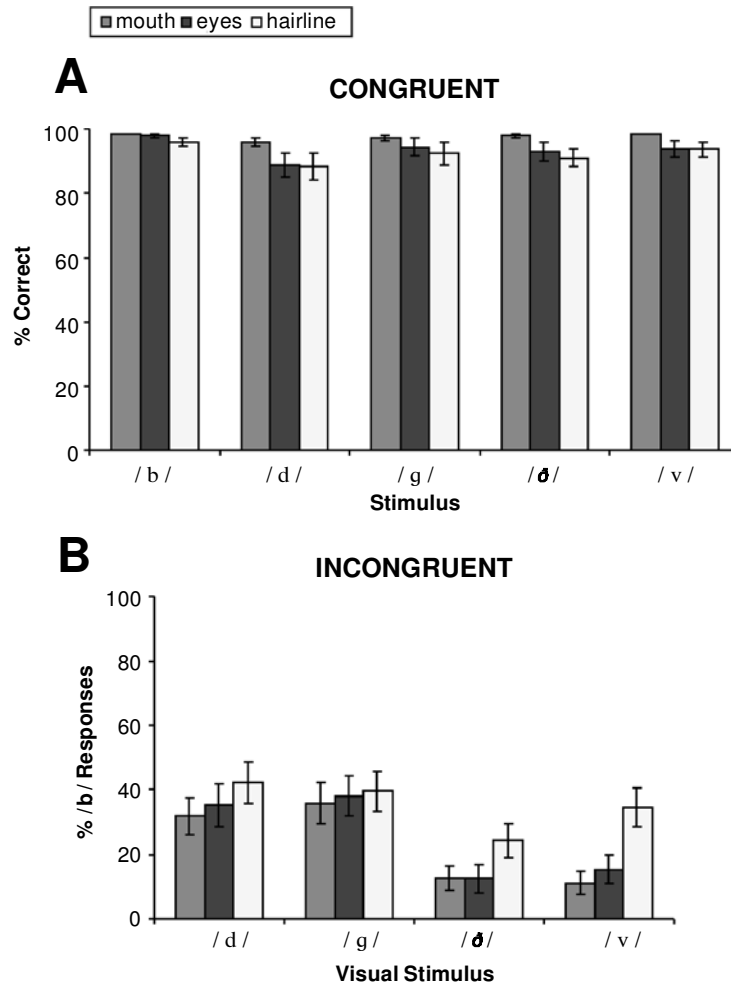
Figure 7. Accuracy performance in response to the presentation of the congruent (A) and incongruent (B) audiovisual stimuli in three different gaze fixation conditions (mouth, eye, and hairline of talker) for the 12 subjects in Experiment 2.

fixation position as factors. The main effects of gaze fixation position [$F(2,22) = 6.2, p < .01$], stimulus [$F(3,33) = 13.3, p < .001$], and talker [$F(3,33) = 19.3, p < .001$] were statistically significant. The percentage of correct /b/ responses was significantly higher in the hairline fixation condition than in either the eye or the mouth fixation condition; there was no significant difference between the responses in the eye and the mouth fixation conditions. For all the incongruent stimuli combined, the average percentage of /b/ responses in the mouth, eye, and hairline fixation conditions varied from 22.9% to 25.3% to 35.2%, respectively. It appears that, in this experiment, not all the stimuli were equal in producing the McGurk effect: There were lower percentages of correct /b/ responses for both /ð/ and /v/ than for /d/ and /g/ (see Massaro, 1998). The same was true for the talkers, with Talker 4 eliciting a significantly weaker McGurk effect than did the other three talkers and Talker 2 eliciting a significantly weaker McGurk

effect than did Talkers 1 and 3. None of the interactions were significant.

These results are consistent with those obtained in Experiment 1 in that they showed no significant differences in the occurrences of the McGurk effect between gaze fixations on the mouth and the eyes. One could therefore conclude that fixating within the central region (ocular and oral regions) of a talker's face provides identical vantage points for processing visual speech information. Only after moving the gaze fixation position further from the talkers' mouths did the effect become significantly, albeit only mildly, weaker. Unfortunately, we cannot dissociate whether the analysis of visual speech information under this fixation condition was achieved by the peripheral retina's viewing the talker's mouth or from the locally correlated motion, even though the latter possibility is less likely than the former.

Whereas the manipulation of gaze fixation position in the present experiment was vertical, a small body of liter-

ature has focused on the effects of horizontally displacing the gaze of subjects relative to speakers, in order to address questions of hemispheric dominance for speechreading. The reported visual field advantages are, however, not fully consistent. Although both left-field (R. Campbell, 1986) and right-field (R. Campbell, de Gelder, & de Haan, 1996) advantages have been observed for static face tasks, a right visual field advantage was reported for silent speechreading (Smeele, Massaro, Cohen, & Sittig, 1998). Two other studies (Baynes, Funnell, & Fowler, 1994; Diesch, 1995) also showed a right visual field advantage for the audiovisual fusion effect provided by incongruent McGurk stimuli such as ours. Most of these studies used small gaze displacements (<5° from central fixation) and did not directly manipulate stimulus eccentricities, with the exception of Smeele and colleagues (Smeele et al., 1998), who reported a reliable decrease in speechreading performance when the subject's gaze was displaced eccentrically (but within 10°) from the talking face. Although different in design, the small decrease in speechreading performance reported in the latter study is consistent with the effects of gaze fixation position on the perception of the McGurk effect that we observed.

EXPERIMENT 3

The results of the two preceding experiments suggest that the McGurk effect can depend on where on a talker's face the subject's gaze is directed, but not on the subjects' natural gaze behavior, which rarely involves gaze fixations beyond the oral and ocular regions of a talker's face. In this last experiment, we sought to determine how the McGurk effect varies with a larger manipulation of the eccentricity of the visual stimulus, by instructing subjects to fixate spots of lights positioned beyond both the talker's head and the video monitor. We reasoned that if the McGurk effect persists under these gaze fixation conditions, any available visual speech information must be analyzed by processes originating within the peripheral retina.

Method

Subjects. Sixteen subjects between the ages of 18 and 20 years were tested.

Stimuli. To reduce the duration of this experiment, we chose to use only three of the four talkers that had been used in the previous experiments. We kept Talkers 1 and 4 because they elicited the two ends of the range of responses. Talkers 2 and 3 generally elicited similar intermediate responses; thus, we chose to drop one of them, Talker 2. A total of 540 trials were tested for each subject—that is, 135 trials for each of four gaze fixation positions: five consecutive blocks with each of the three talkers, each producing nine stimuli, of which five were congruent (/b/, /d/, /g/, /ð/, and /v/) and four incongruent (auditory /b/ dubbed onto visual /d/, /g/, /ð/, or /v/).

Procedure. The subjects were seated 57 cm from the video monitor, with their heads restrained in a head- and chinrest, and viewed the monitor monocularly with their dominant eyes. They were instructed to fixate one of four spatial locations, either the mouth of the talker or a bright spot of light produced by a laser light projected onto a tangent screen, displaced 20°, 40°, or 60° horizontally relative

to the talker's mouth and within the temporal visual field (to avoid the blind spot) of the subject's eye. The order of gaze fixation positions was counterbalanced across subjects. The subjects were instructed to report what they heard on each trial.

Results and Discussion

Figure 8 shows the mean percentage of correct responses for each of the congruent (A) and incongruent (B) stimuli. The data for the incongruent stimulus condition were analyzed with a three-way ANOVA with gaze fixation position, stimulus, and talker as factors. The main effect of gaze fixation position was highly significant [$F(3,45) = 195.2, p < .001$], with the percentage of correct /b/ responses at each gaze fixation position being significantly different from that at any other position. For all the incongruent stimuli combined, the average percentage of /b/ responses increased with the eccentricity of the gaze fixation position, from 23.8% to 44.3% to 80.7% to 88.3%. In comparison, the average percentage of correct responses in the congruent stimulus condition decreased slightly with the eccentricity of the gaze fixation position, from 99.0% to 96.3% to 95.9% to 90.2%, whereas the average percentage of correct responses in the auditory-only condition was 94.6% (Table 1). The main effect of talker was also statistically significant [$F(2,30) = 40.9, p < .001$], with Talker 1 producing a significantly stronger McGurk effect than did both Talkers 3 and 4. No other effect or interactions were found to be statistically significant, except for the interaction between gaze fixation position and talker [$F(6,90) = 2.6, p < .05$]. We examined the significance of the differences in performance between stimulus conditions (congruent vs. incongruent) for each gaze fixation position with a three-way ANOVA with stimulus condition, stimulus, and talker as factors. The difference in the percentage of correct responses between the stimulus conditions was found to be significant at all gaze fixation positions, except when the subject's gaze position deviated 60° from the talker's mouth.

These results extended those obtained in Experiment 2 by demonstrating that visual speech information can still influence auditory perception when it is accessible only via peripheral retina processing: A significant McGurk effect was still produced with the subject's fixation being deviated 20° and even 40° from the talker's mouth. This finding echoes previous observations that have shown how speech perception remains effective when facial images are spatially degraded by quantization (e.g., C. S. Campbell & Massaro, 1997; MacDonald et al., 2000), band-pass filtering (Munhall, Kroos, & Vatikiotis-Bateson, 2003), visual blur (Thorn & Thorn, 1989), or increased stimulus distance (Jordan & Sergeant, 2000). Altogether, these results support the hypothesis that visual speech information is processed at a coarse-scale level, which may be possible because low-resolution information, efficiently processed by the magnocellular pathway originating from the peripheral retina, dominates the early visual information (global image) processing associated with pattern recognition (Hughes, Nozawa, & Kitterle, 1996). Such
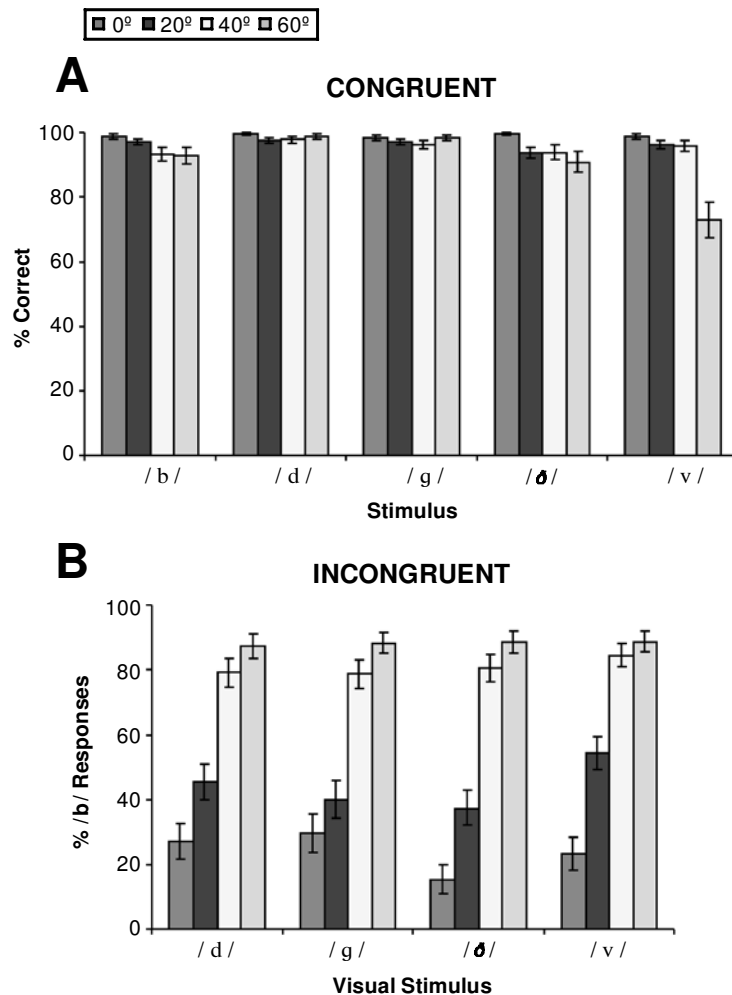
Figure 8. Accuracy performance in response to the presentation of the congruent (A) and incongruent (B) audiovisual stimuli at four different gaze fixation eccentricities relative to the talker's mouth for the 16 subjects in Experiment 3.

rapid visual processes may actually better match in speed the usually faster auditory processes, thereby allowing the integration of information from both modalities, to yield audiovisual speech perception, including the McGurk effect.

Although there was visual speech information available to the subjects when they directed their gaze away from the talker's mouth, its influence on their auditory percept in the incongruent stimulus condition significantly weakened with eccentricity and became nonsignificant when compared with performance in the congruent stimulus condition when gaze deviated 60º from the talker's mouth. This eventual fall of the McGurk effect echoes the observation that the identification of facial images declines with eccentricity (Mäkelä, Näsänen, Rovamo, & Melmoth, 2001). One may assume, however, that the peripheral retina may more effectively process the additional temporal frequency information that dynamic faces contain, relative to static faces. Visual motion could indeed

be an essential factor for detecting the shape deformation produced by facial movements (Loffler & Wilson, 2001), despite the limitations of peripheral motion processing. Whereas some high temporal frequency processing can be relatively invariant with respect to eccentricity, low temporal frequency processing is particularly sensitive to eccentricity (Tynan & Sekuler, 1982). It is therefore tempting to speculate that significant information from facial movements, perhaps from the rapid articulations associated with the production of the consonants, may have been available and may have been integrated with the speech acoustics to produce reliable McGurk effects when the subjects directed their gaze beyond the talker's face.

It is noteworthy to point out that manipulating stimulus eccentricity also reduced the performance in the congruent condition, although the identification of the congruent /b/ stimulus at a large gaze eccentricity (40º; 95.4%) remained better than the corresponding performance in the incongruent stimulus condition (80.7%). Two aspects of

these results are important. First, there was a substantial McGurk effect at such eccentric gaze fixation positions, whereas the congruent /b/ stimulus was only slightly affected. This differential visual influence of stimuli presented at a large eccentricity is consistent with the idea that visual speech needs to provide more information to impair identification of incongruent auditory speech than is needed to improve identification of congruent auditory speech (Massaro, 1998). Second, at such eccentric gaze fixation positions, there was still enough visual information to distinguish the /b/ stimulus from other visual consonants. This may have been due to a fairly gross detection of oral closure, but this is nonetheless remarkable.

GENERAL DISCUSSION

The findings from the three experiments in this study present a consistent picture of visual speech processing: Subjects do not fixate and, more important, need not fixate exclusively on the talker's mouth to perceive linguistic information. In Experiment 1, natural gaze patterns varied across the face, and the location of gaze fixation at the time of consonant production did not predict the likelihood of perceiving the McGurk effect. In Experiments 2 and 3, manipulation of the subjects' gaze fixations within the talker's face did not influence audiovisual speech perception substantially, and it was not until their gaze was displaced beyond 10°–20° from the talker's mouth that the McGurk effect was significantly affected. These results further demonstrate both the strength of the influence that visual speech information exerts on auditory speech perception and the remarkable operational range of visible speech processing.

Gaze in natural communication contexts is driven by social and conversational factors, as well as by the occasional need to enhance intelligibility. People's gaze is drawn to the talker's eyes, to the talker's hands producing manual gestures, and to objects that are being discussed and is even just averted to break off eye-to-eye contact. Our findings indicate that the visual influences on speech processing can be effective over a wide range of these gaze fixations. This is fortuitous for speech perception, since it means that intelligibility can be enhanced even when gaze direction is being determined by other social priorities.

The results of this study are consistent with the observation that visual speech perception can subsist when experimental stimuli are restricted to low spatial frequency components of the images. Video image quality has been manipulated using various techniques that eliminate part of the spatial frequency spectrum. Studies using quantized facial images (e.g., C. S. Campbell & Massaro, 1997; MacDonald et al., 2000) have shown that stimuli containing low spatial frequency information still convey the visual information necessary to produce the McGurk effect. Similarly, Munhall et al. (2003), using band-pass filtered visual stimuli, have shown that the peak gain in intelligibility in a speech-in-noise task occurred for the band with a center frequency of 11 cycles per face but that the band with 5.5 cycles per face also significantly enhanced intelligibility. In addition, speechreading performance has been found to remain relatively unaffected with the introduction of visual blur (as much as 4 diopters) by means of convex lenses (Thorn & Thorn, 1989). Finally, Jordan and Sergeant (2000) showed that the McGurk effect persists even with visual speech information provided by quite distant faces. Altogether, these different findings converged to one conclusion: The high spatial frequency information gained by oral foveation is not necessary for visual speech perception.

Along with our report, the small body of literature on gaze behavior in audiovisual speech perception (Lansing & McConkie, 1999; Vatikiotis-Bateson et al., 1998) consistently has shown that subjects fixate on the salient features of a talker's face during experimental tasks. The distribution of gaze fixations within the face, however, varies slightly depending on the experiment. The instructions and experimental context presumably bias the pattern of gaze behavior, and thus any single experiment must be interpreted within the confines of its task demands. The experimental paradigms used to date, including that in the present study, differ markedly from the information processing required during natural interactions, and thus we still know little about the contingencies of dynamic gaze control during real conversations.

The aim of the present experiments were, however, somewhat different from the question of where people look. We tested what visual information could be perceived as a function of gaze position. Although our data show consistently that parafoveal and peripheral visual information can influence auditory speech processing, a number of influences on these data are unknown. Visual attention was a largely uncontrolled variable in our experiments, our assumption being that it was closely related to gaze direction. This is, however, not necessarily the case (e.g., Posner, 1980; Reeves & Sperling, 1986), although some have argued that when subjects are allowed to search a visual scene, there is very little room for attentional shifts additional to those associated with the execution of eye movements (Maioli, Benaglio, Siri, Sosta, & Cappa, 2001). Nevertheless, the contribution of attention to audiovisual integration needs cannot be determined from our data. In a series of experiments, Massaro (1998) examined whether the strength of the McGurk effect changed if subjects were instructed to focus on one modality or both and found that, relative to the bimodal instruction condition, the effect was attenuated when subjects based their responses on the auditory information and it was enhanced when they based their judgment on the visual information. Although visual information always influenced the auditory percept in Massaro's (1998) experiments, thus producing the McGurk effect, its strength varied with instructions.

In Experiments 2 and 3, the subjects were instructed to fixate on various locations but to watch the centrally presented face. This amounts to dividing attention across two

spatial locations and, as Driver (1996) has shown, auditory speech perception can be influenced by shifts in the spatial location of visual attention. Furthermore, a visually presented secondary task can modulate the influence of some auditory speech cues (Gordon, Eberhardt, & Rueckl, 1993). In our data, attentional factors could have had at least two influences on auditory perception. First, fixating and attending away from the acoustic source could have interfered with auditory processing. This would have reduced the intelligibility of the acoustic /b/ for any of the gaze fixation positions displaced from natural gaze patterns. A second possibility is that the distribution of attention might have varied systematically with the gaze fixation manipulation. Neither of these possibilities could account for the results completely, since they both predict a decrease in the percentage of /b/ responses when gaze fixation is displaced toward the periphery. Admittedly, any number of complex interactions between the available visual information and attention are in principle possible. Further studies are needed to address this issue.

A final unknown in our experiments is the distribution of temporal frequencies of the stimuli we used. The power spectrum for speech movements is dominated by low-frequency components associated with the slow opening and closing of the vocal tract to produce a sequence of syllables. The modal syllable rate is ~4–5 Hz (e.g., Greenberg, 1999), and face movements also show a predominance of low-frequency movements across the entire face (Munhall & Vatikiotis-Bateson, 1998). There are, however, some rapid articulations associated with consonant production. How important such high temporal frequency speech information is to intelligibility remains unclear, but it is possible that these frequency components contribute to the McGurk effect resilience to manipulation of stimulus eccentricity. Unlike the processing of low temporal frequency, which seems to dominate visual speech, high temporal frequency processing can indeed be preserved for relatively eccentric stimuli (Tynan & Sekuler, 1982). The importance of visual motion in audiovisual speech perception is further predicted by its essential role in the detection of shape deformation of moving patterns (Loffler & Wilson, 2001). Future studies involving selective temporal frequency filtering are required in order to understand the exact contribution of facial movements in visual speech processing.

The McGurk effect is viewed as a prototypical example of cross-modal integration, but perception of these stimuli also involves integration of many kinds of information beyond integration across the two sensory modalities. Visual speech information is spatially distributed, with the head and motion across the full face providing linguistic information (Munhall & Vatikiotis-Bateson, 1998). These distributed facial movements provide independent information about the utterance, and subjects integrate visual motion across the facial surface to form the percept. Statistical analyses have shown that different parts of the face contribute independently to predictions of the acoustic structure of an utterance. Seeing more of the face in au-

diovisual speech-in-noise experiments (Benoît, Guiard-Marigny, Le Goff, & Adjoudani, 1996) thus increases intelligibility more than does simply seeing the lips. In addition to this spatial integration, subjects must integrate information over time and across fixations. Our conclusions from Experiment 1 were based on the location of gaze at one instant, whereas the visual information from the face is distributed temporally and information uptake is extended in time (Munhall & Tohkura, 1998). It is therefore worth investigating whether the gathering of visual speech information could be more fully reflected in the scan path taken by the subjects' gaze, even though our present data seem to concur with the idea (see Viviani, 1990) that gaze behavior is rather weakly correlated with perceptual processes.

REFERENCES

Baynes, K., Funnell, M. G., & Fowler, C. A. (1994). Hemispheric contributions to the integration of visual and auditory information in speech perception. *Perception & Psychophysics*, **55**, 633-641.

Benoît, 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. Hennecke (Eds.), Speechreading by humans and machines: Models, systems and applications (pp. 315-328). New York: Springer-Verlag.

Benton, A. (1990). Facial recognition 1990. *Cortex*, **26**, 491-499.

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. *Perception & Psychophysics*, **62**, 233-252.

Burt, M. D., & Perrett, D. I. (1997). Perceptual asymmetries in judgements of facial attractiveness, age, gender, speech and expression. *Neuropsychologia*, **35**, 685-693.

Campbell, C. S., & Massaro, D. W. (1997). Perception of visible speech: Influence of spatial quantization. *Perception*, **26**, 627-644.

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

Campbell, R., de Gelder, B., & de Haan, E. (1996). The lateralization of lip-reading: A second look. *Neuropsychologia*, **34**, 1235-1240.

Collewijn, H., van der Mark, F., & Jansen, T. C. (1975). Precise recording of human eye movements. *Vision Research*, **15**, 447-450.

Demorest, M. E., & Bernstein, L. E. (1992). Sources of variability in speechreading sentences: A generalizability analysis. *Journal of Speech & Hearing Research*, **35**, 876-891.

Diesch, E. (1995). Left and right hemifield advantages of fusions and combinations in audiovisual speech perception. *Quarterly Journal of Experimental Psychology*, **48A**, 320-333.

Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature*, **381**, 66-68.

Gagné, J. P., Masterson, V., Munhall, K. G., Bilida, N., & Querengesser, C. (1995). Across talker variability in speech intelligibility for conversational and clear speech. *Journal of the Academy of Rehabilitative Audiology*, **27**, 135-158.

Gordon, P. C., Eberhardt, J. L., & Rueckl, J. G. (1993). Attentional modulation of the phonetic significance of acoustic cues. *Cognitive Psychology*, **25**, 1-42.

Green, K. P., & Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics*, **45**, 34-42.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics*, **50**, 524-536.

Greenberg, S. (1999). Speaking in shorthand: A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, **29**, 159-176.

Hays, A. V., Richmond, B. J., & Optican, L. M. (1982). A UNIX-based

multiple process system for real-time data acquisition and control. *WESCON Conference Proceedings*, **2**, 1-10.

HUGHES, H. C., NOZAWA, G., & KITTERLE, F. (1996). Global precedence, spatial frequency channels, and the statistics of natural images. *Journal of Cognitive Neuroscience*, **8**, 197-230.

JORDAN, T., & SERGEANT, P. (2000). Effects of distance on visual and audiovisual speech recognition. *Language & Speech*, **43**, 107-124.

LANGTON, S. R., WATT, R. J., & BRUCE, I. I. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, **4**, 50-59.

LANSING, C. R., & MCCONKIE, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language, & Hearing Research*, **42**, 526-538.

LOFFLER, G., & WILSON, H. R. (2001). Detecting shape deformation of moving patterns. *Vision Research*, **41**, 991-1006.

MACDONALD, J., ANDERSEN, S., & BACHMANN, T. (2000). Hearing by eye: How much spatial degradation can be tolerated? *Perception*, **29**, 1155-1168.

MACDONALD, J., & MCGURK, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, **24**, 253-257.

MAIOLI, C., BENAGLIO, I., SIRI, S., SOSTA, K., & CAPPA, S. (2001). The integration of parallel and serial processing mechanisms in visual search: Evidence from eye movement recording. *European Journal of Neuroscience*, **13**, 364-372.

MÄKELÄ, P., NÄSÄNEN, R., ROVAMO, J., & MELMOTH, D. (2001). Identification of facial images in peripheral vision. *Vision Research*, **41**, 599-610.

MASSARO, D. W. (1987). *Speech perception by ear and eye*. Hillsdale, NJ: Erlbaum.

MASSARO, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

MASSARO, D. W., & COHEN, M. M. (1993). Perceiving asynchronous bimodal speech in consonant–vowel and vowel syllables. *Speech Communication*, **13**, 127-134.

MCGURK, H., & MACDONALD, J. W. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748.

MCNEILL, D. (1992). *Hand and mind*. Chicago: University of Chicago Press.

MILLER, G. A., & NICELY, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, **27**, 338-352.

MOSCOVITCH, M., SCULLION, D., & CHRISTIE, D. (1976). Early versus late stages of processing and their relation to functional hemispheric asymmetries in face recognition. *Journal of Experimental Psychology: Human Perception & Performance*, **2**, 401-416.

MUNHALL, K. G., GRIBBLE, P., SACCO, L., & WARD, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, **58**, 351-362.

MUNHALL, K. G., KROOS, C., & VATIKIOTIS-BATESON, E. (2003). *Spatial frequency requirements for audiovisual speech perception*. Manuscript submitted for publication.

MUNHALL, K. G., & TOHKURA, Y. (1998). Audiovisual gating and the time course of speech perception. *Journal of the Acoustical Society of America*, **104**, 530-539.

MUNHALL, K. G., & VATIKIOTIS-BATESON, E. (1998). The moving face during speech communication. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye: Part 2. The psychology of speechreading and auditory–visual speech* (pp. 123-136). Hove, U.K.: Psychology Press.

POSNER, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, **32**, 3-25.

RAYNER K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, **124**, 372-422.

REEVES, A., & SPERLING, G. (1986). Attention gating in short-term visual memory. *Psychological Review*, **93**, 180-206.

REINGOLD, E. M., CHARNESS, N., POMPLUN, M., & STAMPE, D. M. (2001). Visual span in expert chess players: Evidence from eye movements. *Psychological Science*, **12**, 48-55.

RHODES, G. (1985). Lateralized processes in face recognition. *British Journal of Psychology*, **76**, 249-271.

RICHARDSON, C. K., BOWERS, D., BAUER, R. M., HEILMAN, K. M., & LEONARD, C. M. (2000). Digitizing the moving face during dynamic displays of emotion. *Neuropsychologia*, **38**, 1028-1039.

ROBINSON, D. A. (1963). A method of measuring eye movements using a scleral search coil in a magnetic field. *IEEE Transactions in Biomedical Engineering*, **10**, 137-145.

SEKIYAMA, K., & TOHKURA, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America*, **90**, 1797-1805.

SMEELE, P., MASSARO, D., COHEN, M., & SITTIG, A. (1998). Laterality in visual speech perception. *Journal of Experimental Psychology: Human Perception & Psychophysics*, **24**, 1232-1242.

STORK, D. G., & HENNECKE, M. (1996). *Speechreading by humans and machines: Models, systems and applications*. New York: Springer-Verlag.

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

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

TANENHAUS, M. K., MAGNUSON, J. S., DAHAN, D., & CHAMBERS, C. (2000). Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. *Journal of Psycholinguistic Research*, **29**, 557-580.

THORN, F., & THORN, S. (1989). Speechreading with reduced vision: A problem of aging. *Journal of the Optical Society of America*, **6**, 491-499.

TYNAN, P., & SEKULER, R. (1982). Motion processing in peripheral vision: Reaction time and perceived velocity. *Vision Research*, **22**, 61-68.

VATIKIOTIS-BATESON, E., EIGSTI, I.-M., YANO, S., & MUNHALL, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics*, **60**, 926-940.

VATIKIOTIS-BATESON, E., MUNHALL, K. G., HIRAYAMA, M., KASAHARA, Y., & YEHIA, H. (1996). Physiology-based synthesis of audiovisual speech. In *Proceedings of 4th Speech Production Seminar: Models and data* (pp. 241-244). Autrans, France.

VIVIANI, P. (1990). Eye movements in visual search: Cognitive perceptual and motor control aspects. In E. Kowler (Ed.), *Reviews of oculomotor research*: *Eye movements and their role in visual and cognitive processes* (Vol. 4, pp. 353-393). Amsterdam: Elsevier.

YARBUS, A. L. (1967). *Eye movements and vision*. New York: Plenum.

YEHIA, H. C., RUBIN, P. E., & VATIKIOTIS-BATESON, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, **26**, 23-44.

NOTES

1. The use of this dependent measure is consistent with the general definition of the McGurk effect as adopted by others (e.g., Jordan & Sergeant, 2000; MacDonald et al., 2000).

2. The different apparatus and procedure could be responsible for this low percentage of correct responses (strong McGurk effect), as compared with what was found in the other experiments.

3. Although we did not monitor the subjects' gaze in this experiment, preliminary data with the subjects who participated in Experiment 1 indicated that subjects fully comply with these fixation instructions. These pilot data were collected following Experiment 1, and the subjects' compliance led us to carry out Experiments 2 and 3 without the invasive eye monitoring.