Perception & Psychophysics
2003, 65 (4), 536-552

Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences

CHARISSA R. LANSING and GEORGE W. MCCONKIE
University of Illinois at Urbana-Champaign, Champaign, Illinois

In this study, we investigated where people look on talkers' faces as they try to understand what is being said. Sixteen young adults with normal hearing and demonstrated average speechreading proficiency were evaluated under two modality presentation conditions: vision only versus vision plus low-intensity sound. They were scored for the number of words correctly identified from 80 unconnected sentences spoken by two talkers. The results showed two competing tendencies: an eye primacy effect that draws the gaze to the talker's eyes during silence and an information source attraction effect that draws the gaze to the talker's mouth during speech periods. Dynamic shifts occur between eyes and mouth prior to speech onset and following the offset of speech, and saccades tend to be suppressed during speech periods. The degree to which the gaze is drawn to the mouth during speech and the degree to which saccadic activity is suppressed depend on the difficulty of the speech identification task. Under the most difficult modality presentation condition, vision only, accuracy was related to average sentence difficulty and individual proficiency in visual speech perception, but not to the proportion of gaze time directed toward the talker's mouth or toward other parts of the talker's face.

Poets and novelists portray the eyes as windows to the soul, and the significance of gaze has been a theme in classical and popular literature. Early studies have maintained that eye contact and shared or mutual gaze provide critical signals in verbal and nonverbal forms of social interaction among humans and other animals (for a review, see Argyle & Cook, 1976). For example, in conversation, a talker may use gaze or movements in the region of the eyes to signal the end of an utterance, a new conversational turn, intonation, or emphasis or to communicate emotional state or content. From early infancy onward, humans and many species of animals respond to the eyes of others as particular objects of attention—behavior that may be learned or innate. Argyle and Cook proposed that although there may be norms about the frequency and duration of appropriate gaze, the use of gaze in human social behavior is a cultural universal and that the eyes of a talker attract maximum interest.

Because the human retina has only a small region that can resolve high spatial frequencies, people tend to direct their gaze to current or anticipated regions of interest in obtaining information about visual objects and events in the world. Early studies of eye movements show that the perceiver's attention is attracted and held by those elements of a complex object that the perceiver deems important (Yarbus, 1967). Thus, it is of interest that, when looking at a face, people tend to direct their gaze most frequently toward the eyes. The gaze is also often directed toward the mouth (Argyle & Cook, 1976). The eyes and mouth are the most expressive and mobile elements of the face. Whereas the eyes may communicate emotion, attitudes, and other relevant information, quantifiable motion of the jaw, cheeks, and mouth is closely related to the temporal and acoustic characteristics of speech (Yehia, Rubin, & Vatikiotis-Bateson, 1998).

Attention to facial movement is important in speech perception. Human infants learn at an early age to associate facial movements with vocalizations. For example, Kuhl and Meltzoff (1982, 1984) have shown that infants make use of visual cues and can detect ambiguities between images of faces and voices. Studies with adults have demonstrated that information about the temporal characteristics of facial motion may be represented by a sparse distribution of dynamic points that can enhance phonetic perception (Rosenblum, Johnson, & Saldaña, 1996). Furthermore, visual cues can influence (or modify) the perception for speech that is heard. In the laboratory, ambiguities between observable speech articulation and corresponding acoustic signals for certain consonant–vowel combinations result in the perception of a new consonant–vowel combination—for example, visual /ga/ combined with an auditory /ba/ may be perceived as /da/ (McGurk & MacDonald, 1976). Talker-specific characteristics (e.g., detailed information about

This work was supported in part by the National Institute on Deafness and Other Communication Disorders, National Institutes of Health, Grant R29 DC 022050. The authors thank Dawn Tutt for her assistance in data collection. Correspondence concerning this article should be addressed to C. R. Lansing, Department of Speech and Hearing Science, 901 South Sixth Street, University of Illinois at Urbana-Champaign, Champaign, IL 61820 (e-mail: crl@uiuc.edu).

an individual talker's speech articulation) may also contribute to differences in a perceiver's proficiency at discriminating among phonetic units (Kricos & Lesner, 1982). The visual aspects of the talker (e.g., dynamic speech articulation information and physical details about a talker's face) can even affect memory for spoken words (Saldaña, Nygaard, & Pisoni, 1996; Sheffert & Fowler, 1995).

Visual cues have been shown to aid speech perception. In everyday situations, social interactions often occur in cluttered auditory environments in which the speech of others, noise, and reverberant room conditions degrade the quality of auditory information. Consequently, the perceiver must attend to the complex movements of the talker's face and gestures and utilize visual cues to understand speech. Similarly individuals who are hard of hearing or deaf may depend on visual phonetic cues for speech perception in quiet, as well as in adverse, listening conditions. In the laboratory, some deaf individuals who are highly proficient in visual speech perception score as high as 80% words correct on unrelated sentences (Bernstein, Coulter, O'Connell, Eberhardt, & Demorest, 1993; Bernstein, Demorest, Coulter, & O'Connell, 1991). Proficiency in visual speech perception is not well understood but is probably associated with enhanced phonetic perception in some individuals (Bernstein, Demorest, & Tucker, 2000). Phonetic information available through vision may complement that available through audition. For example, cues about the place-of-articulation information that allow a listener to discriminate between /p/, /t/, and /k/ are available through vision but are less distinctive through audition, where they are easily degraded by noise (for a review, see Summerfield, 1987). Visual discrimination between some syllables spoken in isolation, such as /ta/ versus /sa/, is quite poor because the observable gestures associated with their production are visually similar. Nevertheless, visual cues can significantly enhance perceptual accuracy of ambiguous or degraded auditory signals (Miller & Nicely, 1955; Sumby & Pollack, 1954).

Visual cues associated with speech articulation are related to auditory stimuli. The percept when spoken language is processed without sound is that it leaves an *auditory trace*, as if it were heard; this was first referred to as *visual hearing* (Mason, 1943). Results reported by Campbell and Dodd (1980, 1982) revealed a serial position curve containing a recency effect on serial recall for lists of vision-only (silent) spoken words by hearing perceivers. An auditory suffix diminished the recency effect obtained for the silent spoken words. These results were consistent with those reported by Crowder and Morton (1969) for serial recall of auditory-only (heard) words and suggested that the encoding of seen (vision-only) speech has some shared properties with heard (auditory-only) speech. Furthermore, recent evidence from functional cortical imaging has demonstrated that vision-only speech perception by individuals with normal hearing is sufficient to activate the auditory cortex in the absence of auditory speech. This finding is not observed in the analysis of written words and letters or the perception of nonlinguistic facial movements (Calvert et al., 1997). Campbell (1998) suggested that the perceivers' percept that silent speech *is heard* may be related to the patterns of bilateral activation of the cortex revealed in the fMRI data reported by Calvert et al. These data show the expected pattern of recruitment of primarily temporo-parietal areas in the right hemisphere, typically associated with visual processing, as well as of the Brodmann area 41 in the primary auditory cortex. Recruitment of the Brodmann area 41 may be associated with an indirect type of activation unique to *silent* speech perception. The concept of visual hearing is further supported by the relation between the kinematics of the external facial movements of the jaw, cheeks, and mouth and the acoustics of speech (Yehia et al., 1998). For some utterances, visual phonetic cues may precede acoustics by as much as 150–200 msec (Abry, Lallouache, & Cathiard, 1996).

Research has indicated that the mouth is the primary source of phonological cues in visual speech recognition in the absence of an auditory signal and that other facial cues may also be helpful. Marassa and Lansing (1995) and Ijsseldijk (1992) found that information from the lips and mouth region alone is sufficient for word recognition; adding facial motion in other areas did not increase speech perception significantly. Greenberg and Bode (1968) demonstrated improvement in consonant identification when full-face motion was available. However, for present purposes, it is important to note that performance with only the lips visible was 56% in the latter study, whereas full-face visibility raised performance to only 59% accuracy. Preminger, Lin, Payen, and Levitt (1998) also showed the ability of observers to perceive phonemes with the mouth masked, but their data also indicate that masking the mouth produced a substantial drop in performance. Massaro (1998) reported that individuals can discriminate among a small set of test syllables without directly gazing at the mouth of the talker. In summary, blocking the view of the mouth region produces a substantial drop in speech recognition performance, whereas blocking the view of the rest of the face produces a much smaller decrement, although one that is sometimes significant. Vatikiotis-Bateson, Eigsti, Yano, and Munhall (1998) have described the extensive correlations among motions at different facial regions, indicating the potential for using off-mouth facial cues in speech perception, but it appears that the most informative cues are those that are present in the region of the mouth.

Despite the superiority of phonological cues in the mouth region, Vatikiotis-Bateson et al. (1998) found that gaze is directed toward the talker's eyes during conversation when both visual information and auditory information are available. They recorded eye movements of participants listening to and watching videotapes of extended monologues under different levels of masking noise. The noise degraded the auditory information. They reported that noise level affects eye behavior: Greater noise increases the time during which the gaze is directed toward the mouth (from about 37% with no noise to about 56% when noise is so high as to make the speech almost unintelligible) and cuts nearly in half the frequency with which the eyes shift between the talker's mouth and the eyes. Remarkably, under the highest noise level used in their study, their participants still directed their gaze to the talker's eyes almost half the time. From this observation, they suggested that "fine-grained detection of the perioral structures was not necessary for the visual enhancement effect of the stimulus monologues on perception" (p. 936). They further suggested that "it may be better for perceivers not to foveate continuously on the mouth . . . [and that] . . . by foveating primarily on the eyes during audio–visual perception, spatial acuity might be exchanged for the more accurate temporal detection of perioral events afforded nonfoveally" (p. 936). This is supported by the results from an earlier study, in which they concluded that "cues from the lips must be coming from peripheral vision" (Vatikiotis-Bateson, Eigsti, & Yano, 1994a, p. 680).

The purpose of the present study was to further examine where perceivers with some natural speechreading proficiency direct their eyes when trying to understand what a talker is saying under conditions of minimal and no audio information, thus extending the range of conditions used in the Vatikiotis-Bateson (Vatikiotis-Bateson et al., 1994a; Vatikiotis-Bateson et al., 1998) studies. Of particular concern is whether their findings generalize to spoken sentence perception under these conditions. Two explanations for the frequent gazes at the eyes were proposed. One is that perceivers acquire phonetic-related information that is distributed broadly on a talker's face, and the other is that perceivers acquire adequate information from the talker's mouth through peripheral vision, making gazes toward the mouth unnecessary for speech understanding. The latter supposition could be accounted for by Posner's theoretical model that distinguishes overt from covert attention (Posner, 1980; Posner & Raichle, 1994). Overt orienting requires movements of the eyes, but attention shifts can occur covertly in the visual field without any change in eye position. Consequently, a perceiver may examine several locations in the visual field away from the point of fixation covertly, without observable eye movement. Although Posner demonstrated separation between attention and the fovea in laboratory experiments, he contended that this is not a normal property of visual attention.

Still, in speech perception, perceivers probably prefer to utilize auditory information to the extent that it is available and may direct their gaze to the talker's eyes for social reasons. If the speech perception task requires detailed phonetic information, the perceiver may either covertly disengage attention to inspect other facial regions or overtly move the eyes to those regions of the face that convey information thought to be important for successful completion of the task. Posner (1980) observed that when participants are "free to move [their eyes] in an acuity demanding task, they clearly prefer to do so and the different levels of performance with foveal and nonfoveal vision confirm the wisdom of their preference" (p. 9). In visual speech perception, the shift in gaze away from the eyes to orofacial movements could be directed by a central decision that the task requires specific phonetic speech information, or attention may be drawn to the mouth by movements detected in the periphery when gaze is directed to the eyes of the talker. The question to be addressed here is the frequency with which participants, when demonstrably using visual information to facilitate speech perception, direct their gaze toward a talker's eyes, picking up visual cues peripherally (covertly), as opposed to making overt orienting responses that direct the eyes to the mouth.

The Present Study

Participants were selected who demonstrated some natural proficiency in visual speech perception. Our rationale was that if participants did not have the proficiency to use visual cues to understand silent speech to some degree, it was likely that frustration with the task might discourage their search for speech-related information on the talker's face or that facial movement associated with speech production might not be perceived as useful. Furthermore, it seemed plausible that if a perceiver did not experience success in understanding silent speech, this itself could influence visual attention and eye behaviors.

The task chosen for the present study was short-term recall of unrelated spoken utterances presented by a videotaped face either with sound at a low-intensity level or with no sound. The task was designed so as to require close attention to detailed phonetic information produced by the talkers. In addition, the eye movements of the perceivers were monitored while they were attending to the talker in order to obtain a detailed record of the sequence and duration of eye fixations directed to different parts of the talker's face during the video sequence. The eye movement records were time-linked to the frames of the video of the talking face. The spatial and temporal characteristics of the perceivers' eye movements were mapped onto corresponding regions of the talker's face, displayed on a computer monitor. The combination of participants with some speechreading proficiency and isolated sentences allowed us to address a critically important aspect of this type of research: demonstrating that perceivers are actually acquiring and using visual information from the face to facilitate their speech recognition. Unless this is demonstrated, the question of what facial information is being used or from where it is being acquired is moot. With a silent presentation condition, it is obvious that any success in the task is based on the use of visual information. Having speechreaders who demonstrate some natural speechreading proficiency increases the likelihood that they will be able to acquire and use visual facial information. Using isolated sentences reduces language constraint and predictability associated with related-sentence sets or passages, thus requiring greater dependence on the observable speech gestures providing phonetic and lexical information.

Previously recorded sentences spoken by two talkers were chosen for the present study. For one of the talkers, the male, the accuracy of the visual perception of spoken utterances without sound has been shown to be higher than that for the other, the female (Bernstein et al., 1993; Bernstein et al., 1991; Bernstein et al., 2000). It is plausible that these performance differences are related to the availability of visual phonetic cues on the faces of the talkers. However, other talker-specific characteristics, such as physical details about the talker's face or dynamic speech articulation, may contribute to differences in performance as well (Saldaña et al., 1996; Sheffert & Fowler, 1995). Furthermore, it is plausible that performance differences in the accuracy of visual speech perception attributed to talker-specific characteristics may not generalize to performance in visual-plus-auditory speech perception and, thus, should be investigated.

Another consideration in the present study was that of image size. We felt that there are three factors to take into consideration in selecting the size of the face image to use for eye movement research. The first concerns the typical encounter with faces in daily communication. Thus, a normal-sized face at a common distance from a talker would be most appropriate. This would ensure that the facial parts and movement extents are in the normal range, from the perspective of what people commonly experience in the world. The second factor concerns participant performance or preference. Vatikiotis-Bateson et al. (1998) reported that their participants preferred a face size that was larger than normal. As was noted, this may be because it makes the required visual discriminations easier, since critical visual cues are enlarged. The third factor concerns the spatial resolution of the eyetracker used in the research. Larger faces are to be preferred when the eyetracker being used has lower spatial resolution, in order to make the spatial distinctions needed, relative to the stimulus characteristics. Because our research was conducted using a high-resolution eyetracker, image size was not constrained by equipment limitations. For this reason, we opted for an image size and distance that was most representative of everyday situations.

There were certain differences between the present study and that of Vatikiotis-Bateson et al. (1998). First, participants were selected who had demonstrated proficiency in using visual facial cues to understand spoken language, since we were concerned that people without this ability might not continue to seriously attempt to perform the task under difficult circumstances. Vatikiotis-Bateson et al. (1998) reported the occurrence of this phenomenon in a pilot study. Second, the task used was more exacting. The participants were presented with unrelated sentences, rather than with a continuous monologue, thus reducing contextual constraints, and were asked to repeat the exact wording of the sentences presented, rather than taking a multiple-option test. Performance was measured by the number of correctly identified words. This helped overcome what we perceive as a limitation in Vatikiotis-Bateson et al.'s (1998) study. In the high-noise condition, the participants in that study scored only 25%–40% correct. With two-choice questions plus "Did not hear" and "Heard but do not remember" options, this does not provide clear evidence that these participants' performance was above chance level—that is, that they were actually making use of the visual speech information. In an earlier study (Vatikiotis-Bateson, Eigsti, & Yano, 1994b), participants were asked two to five short questions after each monologue. Although the results for the medium-noise condition provided evidence that the participants had used visual information to improve speech perception, it was not clear whether these were multiple-choice questions. Therefore, instead of posing questions to evaluate speech understanding, we created a more exacting task by requiring word identification in the present study—that is, the chance level of identifying a word would be approximately zero. Third, reduced audio information was achieved by lowered intensity, rather than by increased noise. In addition, the present study included a condition with no audio signal, representing the opposite end of the audio availability continuum from the no-noise condition used by Vatikiotis-Bateson et al. (1998).

METHOD

Participants

Sixteen young adults, all graduate or undergraduate students at the University of Illinois at Urbana-Champaign, ranging in age from 18 to 20 years, were paid $6 per hour to participate in the study. Each participant reported having learned English as a first language and had no previous training or coursework in lipreading, phonetics, linguistics, or speech and hearing science. They demonstrated normal visual acuity or visual acuity corrected to 20/30, as measured with a Bausch and Lomb Modified Orthorater, and bilateral pure tone detection thresholds no greater than 25 dB HL ([re ANSI, 1996] at 0.5, 1, 2, and 4 kHz). Every participant demonstrated some proficiency for visual speech perception by achieving a score of ≥30% words correct on a screening measure and ≥25% words correct for sentences presented without auditory information. The participants selected practiced with the eye-monitoring instrumentation to ensure that they were comfortable wearing the apparatus and that an interpretable eye behavior recording could be obtained. Approximately 80 volunteers participated in the screening measures, and the first 16 who qualified and were willing to participate were enrolled in the study.

Materials

Stimulus materials consisted of video recordings of the 100 CID everyday sentences (Davis & Silverman, 1970) spoken by male and female talkers and recorded on high-quality laser video disc by Bernstein and Eberhardt (1986). According to Davis and Silverman, the CID sentences were designed to be representative of "everyday American speech," and specifications were laid down by a working group of the Armed Forces–National Research Council Committee on Hearing and Bio-Acoustics, chaired by Grant Fairbanks. Davis and Silverman (1970) cited the major characteristics as the following:

> 1. The vocabulary is appropriate to adults. 2. The words appear with high frequency in one or more of the well-known word counts of the

> English language. 3. Proper names and proper nouns are not used. 4. Common nonslang idioms and contractions are used freely. 5. Phonetic loading and "tongue-twisting" are avoided. 6. Redundancy is high. 7. The level of abstraction is low. 8. Grammatical structure varies freely. 9. Sentence length varies in the following proportion: 2– 4 words [=] 1, 5–9 words [=] 2, 10–12 words [=] 1. 10. Sentence forms are in the following proportions: declarative [=] 6, imperative [=] 2, rising interrogative [=] 1, and falling interrogative [=] 1. (p. 492)

Bernstein and Eberhardt's (1986) recordings displayed only the face of the talker, which filled the video display. The screening measure consisted of 20 sentences spoken by the male talker. Of these, 10 of the video recordings for the male talker were presented with low-intensity sound, and 10 were presented with no sound. Each perceiver received a new random order of the 20 sentences. Estimates of lipreading proficiency were based on word-correct scores achieved for the 10 sentences in the vision-only presentation. These sentences were selected from the original 20 sentences used for screening purposes by Bernstein et al. (1991), and word-correct scores were very similar to those obtained for over 150 participants enrolled in previous visual speech perception studies at the University of Illinois. The remaining 80 sentences were used in the experiment. These were divided into four lists of 20 sentences each, with each list having approximately the same number of words.

Apparatus

Video sequences were displayed on a 17-in. color monitor (Sony Vivitron) in 640 × 480 pixel format and were played on a laser disc player (Pioneer Laservision Player LD-V7000) interfaced to a personal computer via a Video-Logic DV-5000 audio-visual graphics card and controlled by software developed with Ten Core Authoring Language (Computer Teaching Corporation, 1994). When required, audio output from the graphics card was routed to two loudspeakers (Altec Lansing) positioned at either side of the computer monitor. This computer was used to instruct the participant, to control the display of the video sequences, and to record the participant's responses. An Ethernet link transferred time stamps of video frames and received calibration stimuli from a second computer that was interfaced to the eyetracker.

Eye movements were collected by means of a binocular eye tracker (EyeLink, SR Research) with a temporal resolution of 4 msec (250 samples per second) and high spatial resolution. Output from the eye cameras was analyzed in real time by image-processing software to track and detect the center of a participant's pupil. The eye-to-display distance was approximately 65 cm, with 28 pixels/deg of visual angle. The system used head tracking to compute true eye rotation angles and gaze position resolution over a 20° horizontal and 17° vertical tracking range, thus allowing for moderate head motion. The device could detect 0.1° saccades, and the perceivers could refixate a point with a difference in eye position of 0.2°.

Design and Procedure

The participants were instructed that the purpose of the experiment was to test their proficiency for short-term sentence recognition and recall with two different talkers and that some conditions would include sentences presented with no sound. They participated in two 1-h test sessions, each with a different talker. For a given session, two blocks of 20 sentences were presented: one with sound and one without sound. In the vision-plus-sound condition, the average intensity level was set at 50 dB SPL to represent softly spoken utterances—thus, to encourage attention to visual cues. It is possible that attenuating (or amplifying) an acoustic signal may have the potential to influence the percept of naturalness for presentations of the visual-plus-auditory stimuli. The orders of the variables were all counterbalanced across participants. Half saw the male in the first session, and half saw the female; half had vision-only first in a session, and half had vision-plus-sound first. Sentence list was counterbalanced across these conditions. Sentences within a block were shown in a new random order for each participant. No participant saw a repeat of the same sentence, spoken by both talkers. This defined a two-way repeated measures design with two within-subjects factors, talker (male vs. female) × presentation modality (vision only vs. vision plus sound), and two between-subjects factors, talker order and presentation modality order.

At the start of each block, the eyetracker cameras were positioned, and the participant was seated in a chair for which height and position of viewing angle could be adjusted. The equipment was calibrated by having the participant successively fixate nine target locations. Next, a calibration accuracy validation was performed in which the perceivers were instructed to refixate on the same nine points. We proceeded with the experiment if the average discrepancy between corresponding points was no more than 0.1°, with no discrepancy being greater than 0.5°. Prior to and following every trial, a calibration point was displayed at the center of the display screen, and the software was corrected for any drift in eye position for that location because of head movement (i.e., drift correction calibration). Acceptable eye fixation records for analysis included those for which drift corrections differed by less than 0.75° of visual angle (approximately equal to the height and width of the male talker's central incisor as it appeared on the computer monitor). In our experiments, we obtained a mean drift correction value of 0.38° (*SD* = 0.18°) of visual angle from the beginnings to the ends of trials across all the trials used in the analysis.

After a brief pause, the first frame of the video sequence for the sentence was presented for 1 sec; then the remaining frames were played at normal speed (30 frames/second). The final frame remained on the screen for 1 sec before the display was cleared. This gave a still image of the talker's face for a period of time both before and after the speaking of the sentence. Next, a response screen was displayed that cued the participant to say the sentence aloud. The experimenter typed the response on a keyboard and displayed it on the participant's monitor for verification. The participant was instructed to read the sentence and check that the typing was accurate. No feedback was given to the participant about the correctness of the sentence. The eye tracker recorded eye behavior from the onset of the face on each trial until the participant's response was recorded.

Measures

Response files were checked for obvious spelling or typographical error (e.g., *opin* for *open*). A computer program developed by Bernstein et al. (1991) counted words correct per sentence.

For data analysis purposes, the display of each talker's face was divided into rectangles corresponding to seven regions of the face: forehead, eyes, left cheek, nose, right cheek, mouth, and chin. The full-motion sequences for each sentence were inspected on a frame-by-frame basis. The dimensions of the rectangles were adjusted for each video clip to accurately define the location of the facial regions throughout the corresponding utterance. To analyze the eye movement data stream in relation to observable facial motion, it was necessary to identify video frame numbers corresponding to the onset and offset of facial movements. Two perceivers independently identified the onset and offset of any observable face motion, including movements in the talker's eye regions. Agreement was within two video frames in 98% of the sequences. Disagreements were resolved by repeated viewing in a frame-forward and frame-reverse order and by inspection of changes in pixel distance between a face boundary location and a cursor graphic overlaid on the video image.

Customized data reduction algorithms (Eyelink, SR Research) were applied to the eye data collected for the 80 sentence trials. The algorithms time stamped and identified the *x*,*y* locations and the durations of eye fixations. Analysis included only data from sentences in which the drift correction was less than 0.75°, indicating stable data over the trial. Each fixation period was then labeled according to which of the seven facial regions it was in and whether it occurred before, during, or after facial motion associated with the presented sentence—including the time it occurred, relative to facial motion onset and offset.

RESULTS

Visual Perception Performance

Group results for mean percentages of words correctly identified for the 80 CID sentences as a function of test condition are summarized in Figure 1. As was expected, the average performance scores in the vision-plus-sound conditions were higher (male, $M = 98.1\%$, $SD = 0.5\%$; female, $M = 94.2\%$, $SD = 0.9\%$) than those in the vision-only conditions (male, $M = 34.2\%$, $SD = 6.3\%$; female, $M = 23.5\%$, $SD = 2.3\%$) for both talkers. In the vision-plus-sound conditions, the scores were greater than 90%. Thus, audition contributed greatly to performance accuracy. In the vision-only condition one fourth to one third of the spoken words were identified, which indicated successful use of visual information by the perceivers since, as was noted above, chance performance would essentially be zero.

A univariate and multivariate repeated measures analysis was performed on the mean proportion of words correct in the 80 CID sentences per participant for each condition. The within-subjects factors were talker and presentation condition; the between-subjects factors were talker order (e.g., first exposure to the female vs. the male talker) and order of presentation condition (e.g., first exposure to the vision vs. the vision-plus-sound condition). The analysis was conducted on untransformed proportion scores, as well as arcsine transformed scores (to stabilize error variance; Weiner, 1962), with the same results. For simplicity, all the results are reported for the untransformed scores. Neither the order in which the perceivers saw the two talkers [$F(1,12) = 0.12$, $MS_e = 6.83$], nor the order of the presentation modality [$F(1,12) = 1.57$, $MS_e = 88.60$] had any effect, nor was there a significant interaction between these two factors [$F(1,12) = 0.22$, $MS_e = 12.30$]. However, the scores were significantly higher ($p < .0001$) for the male talker ($M = 66.13\%$, $SD = 32.77\%$) than for the female talker [$M = 58.86\%$, $SD = 36.59\%$; $F(1,12) = 27.6$, $MS_e = 847.5$] and for the vision-plus-sound condition ($M = 96.15\%$, $SD = 9.46\%$) than for the vision-only condition [$M = 28.84\%$, $SD = 3.49\%$; $F(1,12) = 1{,}752.9$, $MS_e = 72{,}475.4$]. No significant two-way or three-way interactions were obtained. The finding that visual perception of sentences produced by the female talker was less accurate than that for the male talker replicated earlier findings (Bernstein et al., 1993; Bernstein et al., 1991; Bernstein et al., 2000), although in a subset of 25 CID sentences Bernstein et al. (2000) found sentence accuracy to be higher for a female talker.

Eye Behavior

General patterns. Perceivers made fewer eye movements during the time the face was moving ($M = 1.32$ fixations per sec, $SD = 0.58$) than during periods before ($M = 5.59$, $SD = 1.07$) and after ($M = 3.54$, $SD = 1.29$), when the face was still. There was more ocular activity prior to speech onset than after [$t(15) = 5.091$, $p < .001$] and more activity in the static-face periods prior to speech (motion onset) or after speech (motion offset) than during speech motion [$t(15) = 17.889$ and $t(15) = 8.684$, $ps < .001$, respectively].

Typically, the perceivers made one or two particularly long fixations, periods with no intervening saccades, as Lansing and McConkie (1994) and Vatikiotis-Bateson et al. (1998) also noted in their studies. These long fixations were usually initiated during the facial motion associated with talking, rather than during the static-face periods that preceded or followed talking. Sometimes the perceivers did not move their eyes during the entire pe-

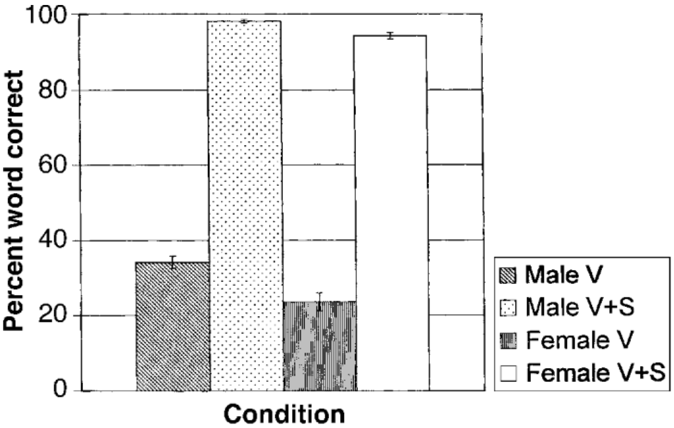

Figure 1. Overall group mean scores (and standard errors of the means) for percentages of words correctly identified for 80 CID sentences. V, vision-only presentation; V+S, vision plus low-intensity sound presentation.

riod that the face was displayed, which will be referred to as *total period* fixations. This occurred on 16.95% of the trials (217/1,280 sentences), 74.65% of which occurred under the vision-only conditions. The total period fixations ranged in duration from 1,212 to 8,052 msec, and in 41% of these the eyes remained in place long after the face had disappeared. Total period fixations were not associated with certain sentences, occurring on 74 of the 80 sentence items (92.50%). Also, these fixations were not limited to a few individuals; all but 2 perceivers demonstrated one or more total period fixations. The average number of total period fixations was 13.56 (*SD* = 11.02). There was wide variability, ranging from 0 to 37 observations per participant (and 0–7 observations per sentence). The 2 participants who made only one or no total period fixations achieved among the lowest scores (25.4% and 27.0% correct) on the speechreading proficiency screening measure; however, there was 1 individual who also achieved a low score (26.7% correct) and used 12 total period fixations. The number of total period fixations per participant in the vision or the vision-plus-sound condition was not correlated with scores on the visual speech perception (speechreading) screening task ($r = -.02$ and $-.2$, ps > .05, respectively).

Distributions of the density of fixations of different durations were examined for the event times (prior to, during, and after facial motion) for each talker/modality-presentation condition. Visual inspection of jitter plots revealed more tightly clustered distributions for the static face conditions that occurred prior to and after the motion than for the face motion conditions. Typically 94.3% and 82.3% of all of the fixation durations were less than 500 msec during the intervals prior to and after face motion, respectively, as compared with only 43.2% of similar duration during the observable motion. In contrast, 34.2% of the gaze durations exceeded 1,000 msec during talking, as compared with 0.3% and 5.2% during the static-face events prior to and after speech production, respectively. Nonparametric statistics were computed for event times (i.e., prior to, during, and after facial motion) within talker/modality condition and are shown in Figure 2. The median values, illustrated by the horizontal line within each box plot, are fairly similar across all events and conditions; however, the number of outliers suggests a tendency for longer fixations to occur during the interval in which there was talking and, possibly, for the interval following talking, as compared with that for the interval prior to talking. These findings support the earlier observations for fixation frequency per second: Fewer and longer gazes occurred during talking than during the periods before and after, when there was no facial motion.

Frequency of eye fixation. An analysis of variance, similar to that used to examine the proportion of words correct, was conducted to test the hypothesis that conditions in which speech recognition is more difficult lead to fewer saccades (and hence, fewer fixations) during speech. The mean number of fixations for the conditions were the following: female–vision, *M* = 52.1 (*SD* = 14.2, *SEM* = 3.5); female–vision-plus-sound, *M* = 79.9 (*SD* = 26.6, *SEM* = 6.7); male–vision, *M* = 48.8 (*SD* = 20.6, *SEM* = 5.2); and male–vision-plus-sound, *M* = 77.3 (*SD* = 32.2, *SEM* = 8.1). This pattern is consistent with regard to the expected modality difference, in that fewer fixations were observed for the vision-alone conditions, on which perceivers achieved lower scores, than for the vision-plus-sound conditions. However the pattern is inconsistent for the expected talker difference, because there were more, rather than fewer, fixations for the female talker than for the male talker. Similarly only the effect of presentation condition was significant [$F(1,12) = 36.4$, $MS_e = 9{,}192.0$, $p < .001$]. Neither the talker nor the talker × presentation condition was significant. Thus, the hypothesis was only partially supported.

Distribution of eye fixation. Two analyses were conducted to determine where people direct their gaze when trying to understand what a talker is saying: one in which the proportion of eye fixations at different facial regions was examined, and a second in which the proportion of the total time that was spent gazing at different regions was examined. These proportions are not independent measures; greater proportions at one facial region occur at the expense of another region. As was indicated above, the talker's face was divided into seven regions. All eye fixation records that met the drift correction criteria were evaluated (12.9% of the records were excluded). Each eye fixation was labeled according to the facial region to which it was directed (e.g., *mouth region*). Any fixation that did not fall within a defined facial region was classified as *out*. Figure 3 presents the percentage of eye fixations directed toward each part of the face while the face was moving or while the face was still, prior to and after the speech, in each of the four experimental conditions. Across full-motion and still-frame segments, of the fixations directed toward the face, more than 86% were directed toward one of three areas: the eyes, the nose, and the mouth. During the still-frame periods, the perceivers' gazes were directed toward the talker's eyes more frequently than toward any other facial region. In contrast, during the full-motion segment, there were more fixations directed toward the mouth than toward any other facial region. The proportion of fixations toward the mouth almost tripled for the speech period over the still-frame period in the vision-only presentation conditions, independently of the talker.

Because so few fixations were directed toward the talker's forehead, cheeks, or chin, the facial area was divided into four major regions: upper face (forehead plus eyes), mid-face (left cheek plus nose plus right cheek), lower face (mouth plus chin), and out (fixation is off of the facial area). Again, only data from trials that met the drift correction criterion were included. For each participant, the amount of time during which the gaze was directed toward each of the four regions while the talker was speaking was calculated, and the time for each region was converted to a proportion of the total gaze time

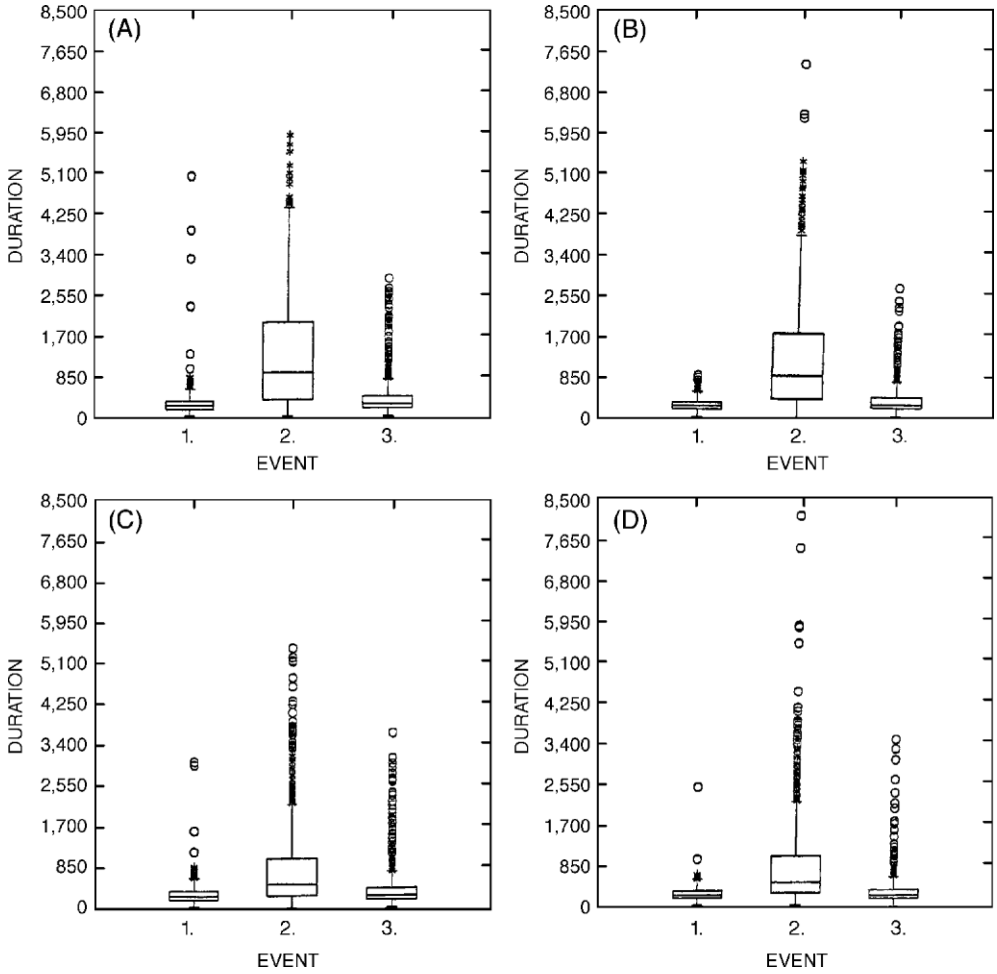

Figure 2. Box plots for the density of eye fixation durations (in milliseconds) under four test conditions as a function of period in which each fixation was initiated: 1 = static face (prior to motion); 2 = face motion (during speech); 3 = static face (after motion). The plots illustrate fixation duration distributions for the following presentation conditions: (A) male talker, vision only; (B) female talker, vision only; (C) male talker, vision plus low-intensity sound; (D) female talker, vision plus low-intensity sound. The bottom and top ends of the boxes indicate the lower (25th percentile) and upper (75th percentile) hinge spread, respectively. The horizontal line across each box indicates the median. The vertical lines extending from the boxes represent the whiskers, and the ends represent the outside values. An outside value is marked with an "*," and a far outside value is marked with an "○."

of facial motion. Means and standard errors for these proportions are shown in Table 1. Across all conditions, the perceivers' eyes were directed toward the lower face (primarily the mouth) for the greatest amount of time. Depending on the talker-plus-presentation condition, the time spent gazing at the mouth was 4 to 67 times as great as the time spent gazing toward the eyes. The proportion of the time spent gazing at the mouth was greater for conditions with no sound than for conditions with sound present, especially for the female talker.

Dynamic aspects of eye behavior. The data presented in the previous section suggest the presence of dynamic shifts of the gaze direction toward and away from the mouth as speech begins and ends. Additional analyses were conducted in order to examine the temporal characteristics of these gaze shifts. This was complicated by the fact that although the median talking time per sentence was 2,442 msec, the actual speaking time varied from sentence to sentence (*M* = 2,482.01 msec, *SD* = 864.41 msec; range, 957–4,554 msec). Thus, it was necessary to select some standard periods of time common to all the trials on which to make comparisons. Two 2-sec periods were used, one from 1 sec prior to the onset of speech until 1 sec after, and the other from 1 sec prior to speech offset until 1 sec after. In some cases in which the speech period was less than 2 sec, these periods over-

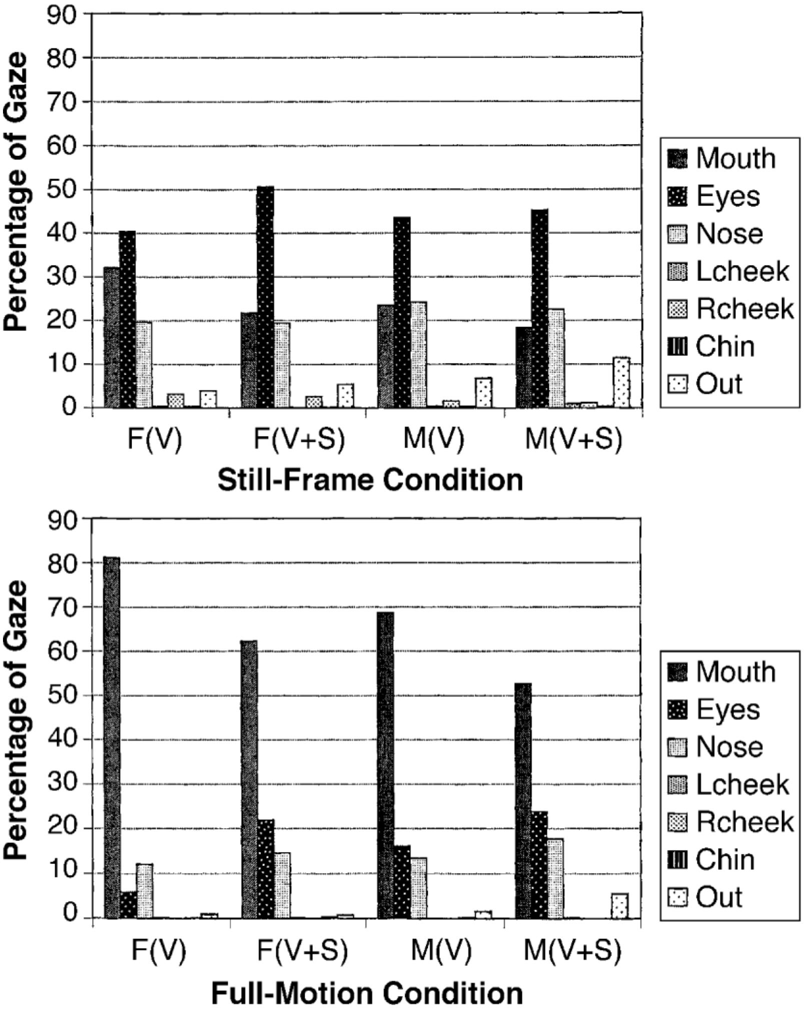

Figure 3. Percentages of eye fixations toward different display regions as a function of facial condition. (Top) Still frame: static face, no motion (pooled data prior to motion onset and after motion offset). (Bottom) Full motion: observable motion associated with speech (prior to and during speech production). Proportion of eye fixation equals fixation count toward a region divided by total number of fixations per talker/modality presentation condition; M(V), male talker, vision only; F(V), female talker, vision only; M(V+S), male talker, vision plus low-intensity sound; F(V+S), female talker, vision plus low-intensity sound.

lapped in time. Data were taken from all trials showing accurate records, as indicated by the drift correction criterion, and for which more than one fixation was made (a single fixation occupied the entire period in 12% of cases, with 83% of these cases showing the gaze to be directed toward the mouth, a fact that will be discussed further below). Figures 4 and 5 show, at each 250-msec interval in the two selected periods, the proportion of trials on which the gaze was directed at the upper (eyes/forehead), middle (nose/cheeks), or lower (mouth/chin) parts of the face or at some other location off of the face (out). In these figures, speech onset and offset are indicated on the x-axis as time zero, with negative values indicating times prior to these events and positive values indicating times after them. The y-axis indicates the proportion of trials on which the eyes were directed at the indicated region at the specified time. Thus, these figures show the dynamic shifts in gaze location that occurred at the times of speech onset and offset. The perceivers began each trial by looking at a drift correction target located at the center of the screen, in what would become the talker's nose region, and the face appeared about 3 sec later. If the gaze did not move during this time, it would be directed at the talker's nose, which was the approximate location of the drift correction target.

As is shown in Figure 4, there were no dramatic shifts of gaze when talking began (the zero point in the time line). In anticipation of the speech, over the 1-sec period prior to speech onset, there was a tendency for people's eyes to

Table 1
Mean Proportions of Times During Facial Motion That Eyes Were Directed to Different Facial Regions

		Display Region							
		Upper		Mid		Lower		Out	
Talker	Modality	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*
F	V	.013	.015	.111	.052	.871	.060	.127	.054
M	V	.114	.059	.125	.034	.795	.073	.047	.014
F	V+S	.103	.036	.096	.040	.796	.061	.015	.005
M	V+S	.164	.051	.180	.049	.646	.074	.066	.045

Note—Proportion of time, total duration of fixations in a region per participant per talker × modality condition/total fixation time per talker × modality condition; upper, forehead + eye regions; mid, right-cheek + nose + left-cheek regions; lower, mouth + chin regions; out, out of the range of the talker's face; F, female talker; M, male talker; V, vision-only presentation; V+S, vision plus sound (approximately 50 dB SPL, soft speech level). Because of drift corrections that exceeded our accuracy criteria, 12.9% of the fixations were discarded and were not analyzed in this experiment. Because data were reported as means for participants, the rows do not necessarily add up to 1.0.

move to the face and, in particular, toward the talker's mouth. However, once the speech began, there was some shifting of the gaze away from the mouth, toward the eyes, or even away from the face. Interestingly, this shift was no more prominent for the vision-plus-speech condition, in which auditory information was available, than for the vision-only condition. Still, the mouth remained the modal gaze location for all except the male talker in the vision-plus-speech condition; in this, the easiest, condition, the modal location eventually became the talker's eyes.

Figure 5, which presents the pattern of dynamic eye behaviors referenced to the offset of facial motion, shows a much more dramatic shift. Although 50%–85% of the fixations were directed toward the mouth region at the end of speaking, within a second this had dropped to 20%, with most of the gaze shifts going quickly to the eyes but with some taking the gaze away from the face. Directing the gaze toward the talker's eyes may have been a learned social response, seeking information about emotional state through facial expression, although it could also have been linked to the anticipation of the written form of the participant's sentence recall response that would soon appear near the top of the display screen, in the region of the talker's forehead. Looking away from the face could have resulted from an attempt to minimize interference in the task of trying to consolidate a response on the basis of the just-received information.

Sentence Identification Accuracy and Gaze Locations

Under every condition, the perceivers directed a larger proportion of their gazes toward the talker's mouth during speech motion than toward other facial regions. This tendency was greater for the more difficult vision-only conditions than for the vision-plus-sound conditions, in which sentence identification accuracy was near perfect. In the vision-only condition, the total number of words in the sentence stimuli and the total number in the response strings were both related to the time that the perceivers gazed at the talker (*r*s = .82 and .41, *p*s < .001, respectively), and it was of interest to investigate whether gaze toward a specific facial region was associated with improved accuracy. It was hypothesized that this identification accuracy would be improved by spending more time gazing at the mouth, as compared with other facial regions. To test this hypothesis, the 83% of the trials (530/640) on which drift corrections were within tolerance limits were identified and submitted to a correlation analysis, to determine whether the proportion of time that the gaze was directed to each specific region on the display of the talker's face was related to the proportion of words reported correctly. Only the correlation between proportion of time on the mouth and accuracy approached statistical significance. It was very low, accounting for less than 1% of the variance, and in the direction opposite the hypothesis: More time on the mouth was associated with poorer accuracy ($r = -.087$, $p = .045$). Other correlations were not significant (i.e., eyes, $r = .075$, $p = .087$; nose, $r = .062$, $p = .155$; and off of the face display region, $r = -.032$, $p = .466$). Thus, the hypothesis was not supported.

The perceivers varied considerably in the proportion of time directed toward any facial region (range on mouth, 44%–99%; range on eyes, 0%–28%; range on nose, 0.2%–44%; range off of the display, 0%–9%). Two perceivers directed their gaze toward the talkers' eyes about 25% of the time and were among the most proficient speechreaders, represented by scores of 45% and 37% correct on the vision-only sentence-screening measures.

Because performance accuracy did not appear to be related much to the proportion of time that the gaze was directed to specific facial regions, particularly the mouth, additional factors were considered. It was plausible that other variables, such as speechreading proficiency or item difficulty, could be hiding the relationship. The participants' word-correct scores on the speechreading screening test were highly correlated with their word-correct scores on the vision-only condition trials ($r = .786$, $p < .002$) and were thus used to represent perceiver proficiency. Word-correct scores across perceivers for

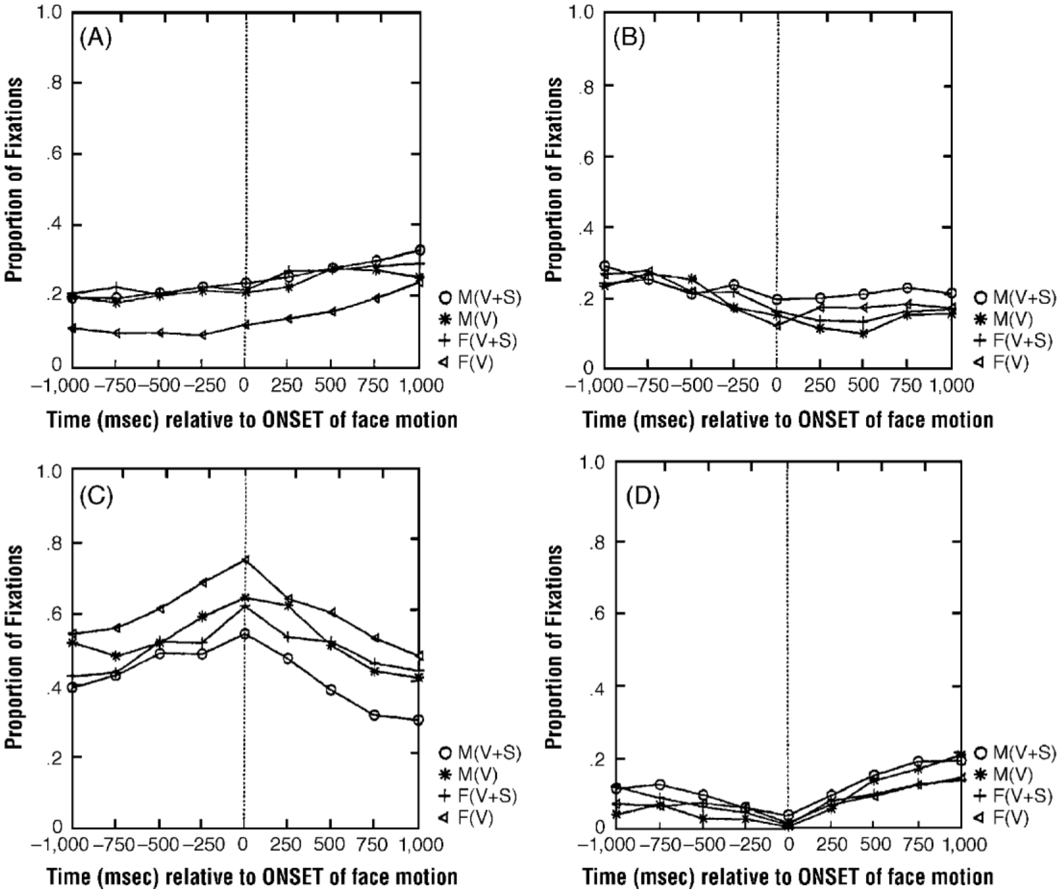

Figure 4. Changes over time in the proportion of eye fixations directed toward different regions of the display for 16 participants in four test conditions referenced to facial motion onset. Plots from A through D show data for the following regions: upper, fixations directed toward the forehead, left-eye, and right-eye regions; middle, fixations directed toward the left-cheek, nose, and right-cheek regions; lower, fixations directed toward the chin and mouth regions; and out, fixations directed off of the face. Frequency of eye fixation is plotted at 250-msec intervals during a 2-sec period, beginning 1 sec before the start of facial motion and ending 1 sec after the start of facial motion. The start-of-facial-motion point is indicated by the dotted vertical line at time = 0. M(V), male talker, vision only; F(V), female talker, vision only; M(V+S), male talker, vision plus low-intensity sound; F(V+S), female talker, vision plus low-intensity sound.

each sentence item were calculated and used to represent item difficulty. Next, correlational analyses were conducted again to incorporate the variables of perceiver proficiency and item difficulty in a multiple regression model. However, partialling out item difficulty ($r = .995$, $p < .001$) and perceiver proficiency ($r = .743$, $p < .001$) did not improve the strength of the correlation between accuracy and frequency of gazes toward the mouth [$r = -.061$; $F(3,636) = 3.941$, $p = .048$], although these variables together accounted for 42.9% of the variance. Similarly, in a separate analysis, partialling out item difficulty ($r = .994$, $p < .001$) and perceiver proficiency ($r = .788$, $p < .001$) did not improve the strength of the correlation between accuracy and frequency of gaze toward the talker's eyes [$r = .051$; $F(3,636) = 1.628$, $p = .203$], although these variables together accounted for 42.5% of the variance. Thus, the failure to support the hypothesis was not due to confounding with these variables.

Overall, the perceivers directed their gaze to more than one facial region on 25.5% of the sentence items in the vision-only conditions. The remaining 74.5% of the items in the vision-only conditions for which eye gaze was limited to a single facial region were identified. On 65% (342/395) of these sentences, the perceivers limited their gaze toward the mouth region, and their average accuracy was 31.8% words correct. In contrast, gaze was limited to the talker's eye region only on 4% (17/395) of the sentences, but the average accuracy was 48.8% words correct. Across all of these items, 100% accuracy was achieved for 45 sentences in which the perceivers limited their gaze to the talker's mouth region, 4 sentences with gaze limited to the eye regions, 6 sentences with gaze limited to the nose region, and 0 sentences with gaze directed off of the face display. However, scores of 0% correct were achieved for 115 sentences in which the perceivers limited gaze to the talker's mouth region, 4 sen-

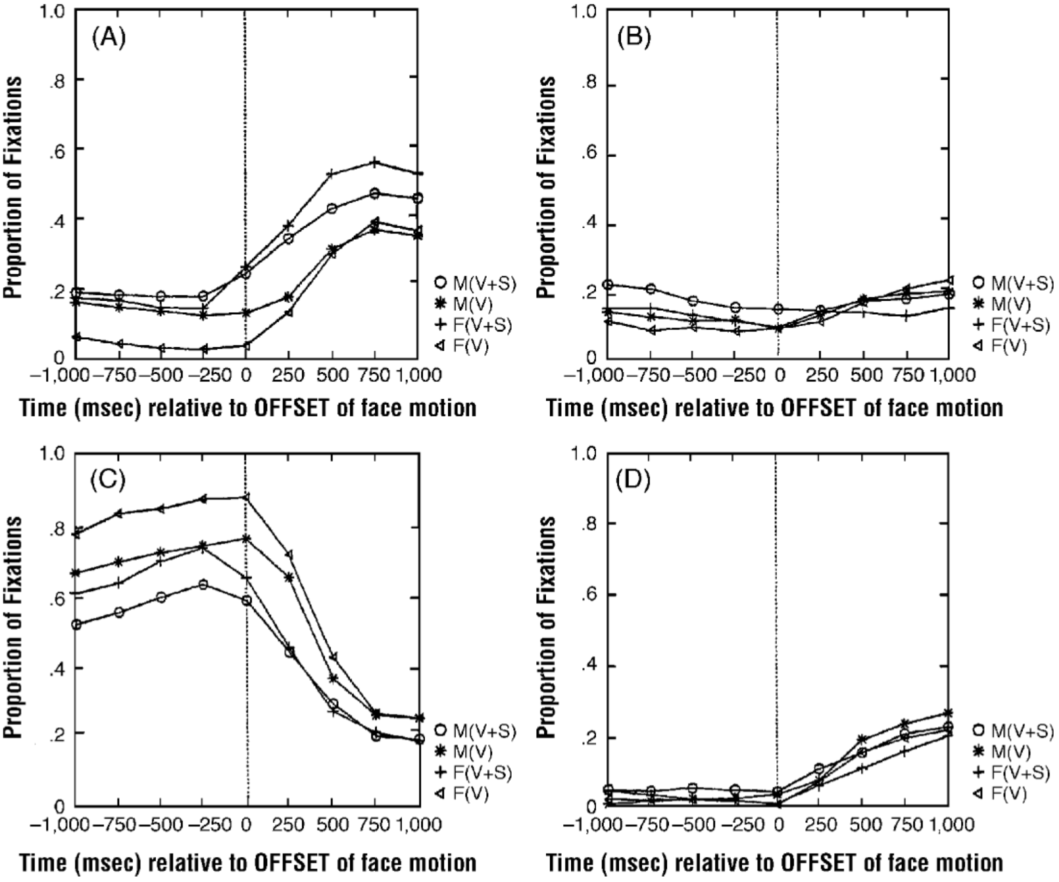

Figure 5. Changes over time in the proportion of eye fixations directed toward different regions of the display for 16 participants in four test conditions referenced to facial motion offset. Plots from A through B show data for the following regions: upper, fixations directed toward the forehead, left-eye, and right-eye regions; middle, fixations directed toward the left-cheek, nose, and right-cheek regions; lower, fixations directed toward the chin and mouth regions; and out, fixations directed off of the face. Frequency of eye fixation is plotted at 250-msec intervals during a 2-sec period, beginning 1 sec before the end of facial motion and ending 1 sec after the end of facial motion. The end of facial motion is indicated by the dotted vertical line at time = 0. M(V), male talker, vision only; F(V), female talker, vision only; M(V+S), male talker, vision plus low-intensity sound; F(V+S), female talker, vision plus low-intensity sound.

tences with gaze limited to the eye regions, 7 sentences with gaze limited to the nose region, and 1 sentence with gaze directed off of the face display. Average performance accuracy by facial region from lowest to highest was the following: off of the display, M = 18.3 (SEM = 16.1); mouth, M = 31.8 (SEM = 1.9); nose, M = 41.6 (SEM = 6.5); and eyes, M = 48.4 (SEM = 9.8). Thus, there was no evidence that spending the full time gazing toward the talker's mouth led to higher performance than did gazing toward other parts of the face.

A final set of correlational analyses was conducted to evaluate whether the number of fixations in specific facial regions was related to sentence accuracy for those sentences in which attention was limited to a single facial region. Partialling out item difficulty (r = .982, p < .001) and perceiver proficiency (r = .698, p < .001) did not increase the strength of the correlation between accuracy and frequency of gazes toward the mouth [r = −.078; F(3,391) = 3.687, p = .056], which accounted for 42.1% of the variance when combined with item difficulty and perceiver proficiency. Similarly, in a separate analysis, partialling out item difficulty (r = .988, p < .001) and perceiver proficiency (r = .645, p < .001) did not improve the strength of the correlation between accuracy and frequency of gaze toward the talker's eyes [r = .092; F(3,391) = 3.368, p = .067], which accounted for 41.6% of the variance when combined with item difficulty and perceiver proficiency.

DISCUSSION

It is well known that people can enhance their perception of speech under nonideal conditions by watching a talker's face. In fact, some people become quite expert at identifying what talkers are saying in the complete absence of an auditory signal. This raises the question of what cues are being used in visual speech perception. The present experiment was conducted to investigate this issue by studying where people direct their gaze as they attempt to identify what a talker is saying under conditions of a low-intensity level versus a completely missing audio signal. The answer is clear: Prior to and after the speech period, the perceiver's eyes are more active, often looking at the talker's eyes (in fact, there are more fixations directed toward the eyes than toward any other part of the face during those periods); however, during the actual speech period there is reduced ocular activity, and the perceiver's eyes are mainly directed toward the talker's mouth. Time plots show dynamic changes in fixation location likelihood. During the second prior to speech onset, there is an increasing proportion of cases in which the eyes are directed toward the talker's mouth, over 85% under some conditions, with an accompanying decrease in fixations at other locations. During the second following speech offset, the eyes tend to move away from the mouth, mainly toward the eyes.

It should be noted that the proportion of fixations toward the nose/cheek area was probably inflated in these data since, prior to the onset of the face, the perceivers had to direct their gaze to a target in the area that would become the nose region when the face appeared, in order to carry out a calibration (drift correction) task. At that point, 100% of the fixations would be on the nose region. By the time the face appeared (indicated as time −1,000 msec in Figure 4), there had already been a shift of fixation locations away from the nose region and toward the regions where the mouth or the eyes would appear; only 25%–40% of the fixations remained in the nose region. Thus, there appear to have been two forces in operation. One draws the eyes to the talker's eyes, possibly for social reasons, as Vatikiotis-Bateson et al. (1998) noted, which we refer to as an *eye primacy effect.* The second force draws the eyes to the mouth when speech-associated motion is taking place or in anticipation of speech, signaled by facial cues or by the audible speech signal, when it is available. We assume that the latter force is the result of attention's being drawn to a location believed by perceivers to be a rich information source, and so we refer to it as an *information source attraction effect.* We suspect that these behaviors are related to attention shifting, described by Posner (1980). In the attempt to understand detailed speech information with an inadequate audio signal, a perceiver often (1) disengages attention from the talker's eyes, (2) moves gaze to a new location (the mouth region) that is perceived as being informative, and (3) engages attention at this new location.

There are three bases for the gaze's being directed to a given region. First, certain stimulus characteristics can attract the eyes, including sudden onsets (Kramer, Cassavaugh, Irwin, Peterson, & Hahn, 2001), local visual motion (Finlay, 1992), and stimulus contrast (Reinagel & Zador, 1999). Second, oculomotor strategies can be developed that produce constrained sequences of saccadic activity (Levy-Schoen, 1981). Third, task considerations can sensitize the perceiver to stimuli having certain characteristics or locations to which the eyes are drawn (Lansing & McConkie, 1999). From this perspective, there are three potential bases for the eye primacy effect. First, the eye region is both a region of high contrast between pupil, iris, sclera, and eyelashes, all contrasting with the surrounding skin, and one of the most active areas of the face in terms of the frequency and degree of motion. Thus, its stimulus characteristics may attract the eyes. Second, people may develop oculomotor strategies, through their experience in speech perception, that favor gazing at the eyes. Third, since the eyes play an important social role in speech communication, often signaling the person being spoken to or the object being spoken about, expressing the talker's emotions, and helping coordinate turn-taking, there may be a learned preference for gazing at the talker's eyes during oral communication. This results from the eyes' role both as an information source and as a source of social-interaction–signaling behavior.

In the present study, the perceiver's eyes tended to be drawn most strongly to the talker's eyes during the periods prior to and after the speech period. These were periods when the face, including the eyes, was fixed. Thus, the eye primacy effect, as observed in this study, was not due entirely to eye motion. Furthermore, this tendency to move toward the eye region following the calibration task began even before the face appeared and, so, was not primarily the result of stimulus contrast. Thus, at least under the present conditions, the eye primacy effect appears to be the result mainly of preestablished tendencies to look to the eyes for strategic or interpersonal communicative and emotion information purposes. Similarly, the mouth-oriented information source attraction effect could be due to the stimulus motion and stimulus contrast that is typically present in that area of the face, to speech-related oculomotor scanning strategies, or to prior experience in knowing that the mouth is a primary source of nonauditory information about phoneme contrasts in spoken language. The fact that in 30%–50% of the cases, the perceiver's eyes shift to the region where the mouth will be, even before the face appears (−1,000 msec in Figure 4), cannot be due to stimulus factors, but only to the perceiver's experience or belief that the mouth region is a source of visual information that will be needed for the task. Once the face appears, there is a further rise in the number of cases in which the perceiver's eyes are directed toward the talker's mouth, even though no mouth motion has yet begun. The shifts of the gaze toward the mouth before and after speech onset and away from the mouth as speech terminates are probably not primarily stimulus driven but are mainly the result of preferences and/or strategies that have been developed through experience. The strength of the mouth-related information source attraction effect and its success at drawing the perceiver's gaze away from other parts of the face, particularly the eyes, depends on the difficulty of the speech perception task—on the degree to which visual information is needed in order to understand the talker's message and the ease with which that information can be obtained. Vatikiotis-Bateson et al. (1998) demonstrated that with noise-free conversational monologues, the perceiver's gaze is directed at the talker's eyes about 60%–65% of the time (as estimated from their Figure 3, p. 930, assuming that most fixations not on the mouth were on the eyes) and that the introduction of increasing amounts of noise reduced this percentage. The present study extended these findings to a situation in which the auditory signal was very weak or entirely missing, showing further increases in eye-directed gaze of up to 85% during actual speech. The value of the auditory signal is seen by the fact that its presence in the present study raised the word-correct score from the 30% range to over 95%. With the auditory signal, visual information was less important (although the present study does not include the data that would be necessary to estimate the relative contributions of visual and auditory information in this condition), and this resulted in a reduction of the time during which the perceivers' gaze was directed toward the talker's mouth.

The amount of time for which the gaze was directed toward the mouth also varied between talkers in the present study, being higher for the female. Past research has indicated that the female talker is, on average, more difficult to understand (Bernstein et al., 1993; Bernstein et al., 1991). The percentages of word-correct scores in the present study were lower for the female talker than for the male, both with the auditory signal present and with it absent. The eye movement data indicate that the perceivers spent more time looking at the female talker's mouth than at the male's, both before and during the spoken message, with the difference being reduced following the end of the message. This difference between the talkers when there was no auditory signal cannot have been due to a difference in the need for visual information in order to carry out the task; in the absence of an auditory signal, the perceivers had to depend entirely on visual information. Thus, the ease with which necessary visual information can be acquired also affects the amount of time that perceivers direct their fixation toward a talker's mouth. With a difficult talker, no auditory information, and a task with little contextual constraint (unconnected sentences) and the need to identify every word, the proportion of time spent with the fixation directed toward the mouth increased to over 85%. The combined results of the present study and that by Vatikiotis-Bateson et al. (1998) indicate a strong positive relation between the difficulty of the speech perception task and the amount of time a perceiver's gaze is directed toward a talker's mouth and, hence, away from a talker's eyes. Thus, the mouth-oriented information source attraction effect increases as the need for and difficulty of obtaining visual speech information increase. The fact that this is a speech-related effect is clearly shown in the analysis of gaze direction over time: As speech onset nears and begins, there is a strong tendency to shift the gaze away from the eyes and toward the mouth, and as the speech ends, the gaze is typically shifted back to the eyes.

It is reasonable to assume that mouth-oriented gazes are a result of past experience in which speech perception under low auditory noise conditions was enhanced by using visual information available from the mouth region; that is, that this information source attraction effect is the result of its past success. On this assumption, one would predict that speech recognition should be improved by directing the gaze toward a talker's mouth, rather than toward the eyes. However, several tests of this prediction in the present study failed to support it: When gazing only at the eyes (eye-only gazes occurred on only 17/395, or 4%, of the sentences presented with no audio signal), the perceivers correctly identified all the words of a sentence in four cases and identified at least one word correctly in another nine cases. Furthermore, there was no increase in the percentage or the number of words identified with increases in the proportion of the sentence-speaking time in which the perceiver's gaze was on the mouth. The sample sizes for these tests were small, owing to the strong tendency to gaze at the talker's mouth under these conditions, and further study is needed. But the present data consistently failed to produce positive evidence for the prediction.

The fact that the perceivers occasionally directed their gaze at the talker's eyes but were able to identify some words supports an observation by Massaro (1998) that visual speech cues can be acquired from peripheral vision. As Vatikiotis-Bateson et al. (1998) pointed out, this may occur in the form either of peripheral cues from the mouth region or of cues from other facial locations. Although Preminger et al. (1998) showed that speech-related cues are available in mid- or upper-facial regions, allowing a perceiver to identify phonetic information accurately in highly constrained speech recognition tasks, in fact the most informative off-mouth areas are the chin and the cheeks. Vatikiotis-Bateson et al. (1998) stated that "it may be better for perceivers not to foveate continuously on the mouth" (p. 936), suggesting that nonfoveal acquisition of motion cues from the mouth or the face is better than foveal acquisition, because of the sensitivity of the visual periphery to temporal information. However, these authors also note that "this account of extracting phonetic information parafoveally could be undermined if subsequent studies showed that subjects adapt to the presence of masking noise by further increasing the proportion of time spent gazing at the talker's mouth" (p. 936). Although the present study did not use masking noise, the same goal was achieved by reducing and eliminating the audio signal, and this did greatly increase perceivers' frequency of directing their gaze at the talker's mouth, suggesting that peripheral acquisition may not be optimal.

The observations from this study raise three questions. Why do people direct their gaze to the mouth if this is not required for successful speechreading? How is visual attention deployed during speechreading? What aspects of processing during speechreading are revealed in eye movement recording?

Given the ample evidence that the mouth region is the primary source of visual speech information, although additional and correlated information is available elsewhere on the face, it is surprising that the present data showed no evidence of speechreading improvement as people directed their gaze at the talker's mouth, rather than at the eyes. Assuming that this finding is replicated in further work, it does not seem to fit comfortably with the observation that, as the speechreading task becomes more difficult, the perceiver's eyes increasingly move to the mouth. If this information source attraction effect does not arise from past reinforcement from greater success under these conditions, why does it exist? One possibility is that people have observed from infancy that the mouth is the actual source of speech; thus, when speech understanding becomes more difficult, there may be a tendency to give attention to the known source of the speech, directing the eyes to that region even though the critical motion is not so fine that it cannot be acquired while the eyes are directed elsewhere on the face. Thus, further research is needed to clarify whether this information source attraction effect is part of learning to optimize speech recognition ability or is the result of a tendency to directly attend the information source even though this is not performance enhancing.

When speechreading is successful with the gaze directed toward the eyes, the relevant facial motion is primarily being acquired peripherally. As was indicated above, present evidence indicates that most of the needed information for successful speechreading must come from the mouth region; eliminating visual cues from the mouth greatly reduces speechreading success. Thus, we assume that successful speechreading involves, to a great extent, success in picking up cues from the mouth. The question needing investigation, then, is whether perceivers direct their attention primarily to the mouth, as the primary information source, or are in fact monitoring the full motion configuration taking place over the face. This is the question of how broadly attention is being distributed during speechreading and whether this varies in a dynamic fashion in real time as information needs change. The eye movements themselves do not suggest dynamic shifts in attention, as reflected by frequent gaze shifts with speech (Vatikiotis-Bateson et al., 1998). Rather, during actual speech periods, there is a tendency to reduce saccades and to hold the gaze in place. It might be argued that attention shifts are being reflected in very small saccades that are not detected by the eyetrackers (Bridgeman & Palca, 1980; Cunitz & Steinman, 1969). That is not likely to be occurring in the present study, since the eyetracker used is very sensitive and saccades longer than 0.1° are reliably detected. However, the eye movements by themselves do not directly reveal information about the size of the region being attended, whether broad or narrow, or about whether covert attention shifts are taking place. These are issues that must be investigated using more sophisticated research methods, probably involving stimulus probes and manipulations, together with eye movement recording.

Thus, there is a need to further explore the nature of eye movement control within the speechreading context, in order to determine how eye behavior relates to attentional and information acquisition activities, as well as social and communicative influences. Findlay and Walker (1999) have proposed a framework for understanding eye movement control from a neurophysiological basis, including the basis for influences from multiple levels of brain activity. The onset time of a saccadic movement is determined by the balance of activity in a fixate center, which suppresses saccadic movement, and a saccade center, which produces saccades. Stimulus motion or attention to stimuli near the center of vision increases activity in the fixate center, whereas stimulus motion in, or the direction of attention to, peripheral visual areas increases activity at the corresponding location in the saccade center. A saccade is initiated when the fixate center activity falls below a threshold, and its spatial characteristics are determined by the region of greatest activity in the saccade center. Yang and McConkie (2001) examined eye behavior during reading and observed how processing difficulties of different types generate inhibition that occurs at different times following fixation onset. From this perspective, the long fixations of the speechreaders may result from a combination of stimulus motion and higher level activity in visual hearing that directs attention to the mouth area. Both of these influences stimulate activity in the saccade center when the mouth is in peripheral vision, drawing the eyes to that location, or stimulate activity in the fixate center when the eyes are already directed toward the mouth area, holding the eyes at that location. We assume that the greater attentional intensity that likely occurs in a task with increased requirements for visual information from a restricted region increases further the activity in the fixate center, further reducing the likelihood that a saccade will be made. The long fixations, then, are cases in which the level of activity in the fixate center is sufficiently high to remain above threshold despite activity, such as that from movement of the talker's eyes or other facial regions, that might simultaneously be occurring in the saccade center. In almost all cases, this occurs when the gaze is directed at the talker's mouth. Differences in this attentional intensity are also seen in the difference in saccadic activity in pre- and postspeech periods and in the speech periods themselves.

From this perspective, the face is a stimulus configuration containing multiple locations where stimulus variation and motion produce activation in the fixate and movement centers, depending on the current fixation location. Previously learned tendencies to seek or strategies for seeking information from different areas under different conditions affect the activation levels in these centers. For example, Althoff and Cohen (1999) and Althoff et al. (1999) indicated that the degree of constraint in the eye movement sequence is greater when a perceiver is looking at a picture of the face of an unknown person than when looking at a picture of a familiar person. Yang and McConkie (2001) included a scanning pattern as a basic component in the control of eye movements in reading. This saccadic activity results from the momentary balances between the activation levels of the move center and the fixate center, which are being influenced by strategic, repetitive activation and by processing activities at other levels. In reading, the primary momentary influence of higher cognitive processes, as observed by Yang and McConkie, is through inhibition that delays saccade onsets and changes the location to which the eyes are sent, thus modifying the decisions that would have occurred on the basis of learned strategies alone. It appears that in speechreading, the attempt to obtain the type of detailed visual data that is required by a motivated person in the absence of a speech signal tends to draw the eyes to the mouth region and produces varying degrees of activation of the fixate center, depending on the task requirements, which is observed as a suppression of the saccades to corresponding degrees. This occurs whether or not moving the eyes to the mouth actually improves speechreading performance. Of course, whether saccades actually occur is also determined by the degree of activation of the saccade center by stimulus and processing events.

REFERENCES

Abry, C., Lallouache, M.-T., & Cathiard, M.-A. (1996). How can coarticulation models account for speech sensitivity in audio-visual desynchronization? In D. Stork & M. Henneke (Eds.), *Speechreading by humans and machines: Models, systems, and applications* (NATO ASI Series, Vol. 150, Series F: Computer and Systems Sciences, pp. 247-255). Berlin: Springer-Verlag.

Althoff, R. R., & Cohen, N. J. (1999). Eye-movement-based memory effect: A reprocessing effect in face perception. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **25**, 997-1010.

Althoff, R. R., Cohen, N. J., McConkie, G. W., Wasserman, S., Maciukenas, M., Azen, R., & Romine, L. (1999). Eye movement-based memory assessment. In W. Becker, H. Deubel, & T. Mergner (Eds.), *Current oculomotor research: Physiological and psychological aspects* (pp. 292-302). New York: Plenum.

American National Standards Institute (1996). *Specifications for audiometers* (ANSI S36-1996). New York: Author.

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.

Bernstein, L. E., Coulter, D. C., O'Connell, M. P., Eberhardt, S. P., & Demorest, M. E. (1993). Vibrotactile and haptic speech codes. In A. Risberg, S. Felicetti, G. Plant, & K.-E. Spens (Eds.), *Proceedings of the Second International Conference on Tactile Aids, Hearing Aids, and Cochlear Implants* (pp. 57-70). Stockholm: Kungliga Tekniska Högskolan.

Bernstein, L. E., Demorest, M. E., Coulter, D. C., & O'Connell, M. P. (1991). Lipreading sentences with vibrotactile vocoders: Performance of normal-hearing and hearing-impaired subjects. *Journal of the Acoustical Society of America*, **95**, 3617-3622.

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

Bernstein, L. E., & Eberhardt, S. P. (1986). *Johns Hopkins Lipreading Corpus I–II: Disc I* [Laser video disc]. Baltimore: Johns Hopkins University.

Bridgeman, B., & Palca, J. (1980). The role of microsaccades in high acuity observational tasks. *Vision Research*, **20**, 813-817.

Calvert, G., Bullmore, E., Brammer, M., Campbell, R., Woodruff, P., McGuire, P., Williams, S., Iversen, S. D., & David, A. S. (1997). Activation of auditory cortex during silent speechreading. *Science*, **276**, 593-596.

Campbell, R. (1998). Everyday speechreading: Understanding seen speech in action. *Scandinavian Journal of Psychology*, **39**, 163-167.

Campbell, R., & Dodd, B. (1980). Hearing by eye. *Quarterly Journal of Experimental Psychology*, **32**, 85-99.

Campbell, R., & Dodd, B. (1982). Some suffix effects on lipread lists. *Canadian Journal of Psychology*, **36**, 509-515.

Computer Teaching Corporation (1994). *TenCore LAS* (Language authoring system: Operating system manual, Version 5.2) [Computer software]. Champaign, IL: Author.

Crowder, R. G., & Morton, J. (1969). Precategorical acoustic storage (PAS). *Perception & Psychophysics*, **5**, 365-373.

Cunitz, R. J., & Steinman, R. M. (1969). Comparison of saccadic eye movements during fixation and reading. *Vision Research*, **9**, 683-693.

Davis, H., & Silverman, S. R. (1970). *Hearing and deafness*. New York: Holt, Rinehart & Winston.

Findlay, J. M., & Walker, R. (1999). A model of saccade generation based on parallel processing and competitive inhibition. *Behavioral & Brain Sciences*, **22**, 661-721.

Finlay, D. (1992). Motion perception in the peripheral visual field. *Perception*, **11**, 457-472.

Greenberg, H. J., & Bode, D. L. (1968). Visual discrimination of consonants. *Journal of Speech & Hearing Research*, **11**, 466-471.

IJsseldijk, F. J. (1992). Speechreading performance under different conditions of video image, repetition, and speech rate. *Journal of Speech & Hearing Research*, **35**, 466-477.

Kramer, A. F., Cassavaugh, N. D., Irwin, D. E., Peterson, M. S., & Hahn, S. (2001). Influence of single and multiple onset distractors on visual search for singleton targets. *Perception & Psychophysics*, **63**, 952-968.

Kricos, P. B., & Lesner, S. A. (1982). Differences in visual intelligibility across talkers. *Volta Review*, **84**, 219-225.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, **218**, 1138-1141.

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior & Development*, **7**, 361-381.

Lansing, C. R., & McConkie, G. W. (1994). A new method for speechreading research. *Journal of the Academy of Rehabilitative Audiology*, **27**, 25-43.

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**, 526-539.

Levy-Schoen, A. (1981). Flexible and/or rigid control of oculomotor scanning behavior. In D. F. Fischer, R. A. Monty, & J. W. Senders (Eds.), *Eye movements: Cognition and visual perception* (pp. 299-318). Hillsdale, NJ: Erlbaum.

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

Mason, M. K. (1943). A cinematographic technique for testing visual speech comprehension. *Journal of Speech Disorders*, **8**, 271-278.

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

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

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, **72**, 338-352.

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

Posner, M. I., & Raichle, M. E. (1994). *Images of mind*. New York: Freeman.

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

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network-Computation in Neural Systems*, **10**, 341-350.

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

Saldaña, H. M., Nygaard, L. C., & Pisoni, D. P. (1996). Episodic encoding of visual speaker attributes and recognition memory for spoken words. In D. Stork & M. Henneke (Eds.), *Speechreading by humans and machines: Models, systems, and applications* (NATO ASI Series, Vol. 150, Series F: Computer and Systems Sciences, pp. 275-281). Berlin: Springer-Verlag.

Sheffert, S. M., & Fowler, C. A. (1995). The effects of voice and visible speaker change on memory for spoken words. *Journal of Memory & Language*, **34**, 665-685.

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

Summerfield, 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 lipreading* (pp. 3-52). Hillsdale, NJ: Erlbaum.

Vatikiotis-Bateson, E., Eigsti, I.-M., & Yano, S. (1994a). Listener eye movement behavior during audiovisual perception. *Proceedings of the Acoustical Society of Japan*, **94-3**, 679-680.

Vatikiotis-Bateson, E., Eigsti, I.-M., & Yano, S. (1994b). Listener eye movement behavior during audiovisual speech perception. In *Proceedings of ICSLP 94: International Conference on Spoken Language Processing* (Vol. 2, pp. 527-530). Tokyo: Acoustical Society of Japan.

Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., & Munhall, K. (1998). Eye movements of perceivers during audiovisual speech perception. *Perception & Psychophysics*, **60**, 926-940.

Weiner, B. J. (1962). *Statistical principles in experimental design*. New York: McGraw-Hill.

Yang, S.-N., & McConkie, G. W. (2001). Eye movements during reading: A theory of saccade initiation times. *Vision Research*, **41**, 3567-3585.

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

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, **26**, 23-43.

(Manuscript received May 7, 2001;
revision accepted for publication December 12, 2002.)