Perception & Psychophysics
1998, 60 (6), 926-940

Eye movement of perceivers during audiovisual speech perception

ERIC VATIKIOTIS-BATESON
ATR Human Information Processing Research Laboratories, Kyoto, Japan

INGE-MARIE EIGSTI
University of Rochester, Rochester, New York

SUMIO YANO
NHK Research Laboratories, Kinuta, Japan

and

KEVIN G. MUNHALL
Queen's University, Kingston, Ontario, Canada

Perceiver eye movements were recorded during audiovisual presentations of extended monologues. Monologues were presented at different image sizes and with different levels of acoustic masking noise. Two clear targets of gaze fixation were identified, the eyes and the mouth. Regardless of image size, perceivers of both Japanese and English gazed more at the mouth as masking noise levels increased. However, even at the highest noise levels and largest image sizes, subjects gazed at the mouth only about half the time. For the eye target, perceivers typically gazed at one eye more than the other, and the tendency became stronger at higher noise levels. English perceivers displayed more variety of gaze-sequence patterns (e.g., left eye to mouth to left eye to right eye) and persisted in using them at higher noise levels than did Japanese perceivers. No segment-level correlations were found between perceiver eye motions and phoneme identity of the stimuli.

It is well known that visual information from the face can influence the perception of speech. Examples of visual enhancement are the ability to "read lips" (Gailey, 1987; Jeffers & Barley, 1971) and the greater intelligibility of speech produced in noise when the speaker's face is visible (e.g., Summerfield, 1987). Somewhat different phenomena are manipulations resulting in the "fusion illusion" of the McGurk effect (McGurk & MacDonald, 1976; see also Green & Kuhl, 1989, 1991; Massaro, 1987; Munhall, Gribble, Sacco, & Ward, 1996) and the ventriloquist effect (e.g., Bertelson & Radeau, 1976). In the McGurk effect, mismatched visual and acoustic events are integrated even when the acoustic signal is clear and perceptible. These result in perceptual shifts such as when auditory /ba/ and visual /ga/ are perceived audiovisually as /da/. The ventriloquist effect requires that the perceiver integrate spatially disparate acoustic and visual events. Perceivers can do this even when the acoustics are masked by distractor acoustics generated at the point of visual origin (Driver, 1996).

It is also known that the visual enhancement of speech perception depends on both dynamic and static characteristics of facial images. For example, Vitkovitch and Barber (1994) have demonstrated that the enhancement effect of visual information deteriorates rapidly as video frame rates fall below about 16 Hz. Furthermore, the temporal characteristics of facial motion enhance phonetic perception even when spatial information is very sparse, as demonstrated by use of dynamic point-light displays (e.g., Rosenblum, Johnson, & Saldaña, 1996; Smeele, 1996).

From the speechreading literature, one might assume that the relevant visual events for speech perception are located at the mouth (e.g., Jeffers & Barley, 1971). Indeed, most engineering efforts to enhance acoustic speech recognition systems with visual features have restricted their search space to the area of the lips and oral aperture (e.g., Benoît, Lallouache, Mohamadi, & Abry, 1992; Brooke & Summerfield, 1983; Luettin, Thacker, & Beet, 1996; Petajan, 1985; Wolff, Prasad, Stork, & Hennecke, 1994). A second major assumption underlying most speechreading research has been that, for both human and machine viewers, the sought-after visual parameters should be extracted with as much precision as possible (e.g., Benoît et al., 1992; cf. Rosenblum et al., 1996, for the use

The authors thank Carol Fowler, James Jenkins, Robert Port, Philip Rubin, and Yoh'ichi Tohkura for helpful criticism and support, Frederique Garcia and Shigeru Mukaida for software support of the binning and gaze sequence analyses, and Hans Tillmann, Philip Hoole, and the Institut für Phonetik und Sprachliche Kommunikation der Universität München for generous support and discussion during a phase of the writing. Christian Benoît, Robert Remez, Jean Vroomen, and an anonymous reviewer provided invaluable assistance with the revisions. Correspondence should be addressed to E. Vatikiotis-Bateson, ATR HIPRL, 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan (e-mail: bateson@hip.atr.co.jp).

of point-light displays). Both of these assumptions lead to the prediction that perceivers should keep the mouth region in the fovea as much as possible.

Why then do people so often report that they “watch” the speaker’s eyes during face-to-face conversation? They are referring to situations that presumably involve more complex linguistic and social behavior than what is encountered in a typical perception experiment. Still, it is curious to assume that perceivers extract precise parameter information from a region of the face toward which they may not foveate or even attend. Of course, perceiver introspection may be wrong. Alternatively, audiovisual perception may be structured in such a way that precise attention to perioral structures does not contradict the subjective impression that the eyes are what perceivers primarily watch.

From such considerations, a number of questions arise about how perceivers extract phonetically relevant visual information from the time-varying behavior of speakers who can be both seen and heard. Perhaps perceivers watch both the eyes and the mouth; but, if so, how—simultaneously or sequentially? To what extent do perceivers “track” phonetic events visually, and in what temporal domain? Are the phonetically relevant visible structures only those in the vicinity of the mouth, such as the lips, tongue tip, and teeth, or is the relevant information less direct and perhaps distributed over larger regions of the face?

In recent examinations of orofacial motion during the production of utterances ranging between repetitive nonsense (e.g., /apaw ... apaw ... /) and spontaneous sentences, it has been shown that lip shape and motion information is distributed over large regions of the face and that acoustic correlates from remote regions of the face are not identical to those provided by the shape and motion of the lips (Vatikiotis-Bateson & Yehia, 1996). That such information is available to perceivers is no guarantee that they actually use it. The purpose of this study was to address these questions from the perceiver’s point of view by examining the one cogent piece of motor behavior exhibited by perceivers during audiovisual speech perception—namely, the perceiver’s eye movements. By examining the kinematics of eye motion and the location(s) of gaze fixation, we may be able to characterize the relevant behavioral patterns and their susceptibility to linguistic and contextual factors in the audiovisual environment.

Thus far, perceiver eye-movement behavior during audiovisual perception tasks has received little attention. To our knowledge, there has been only one published paper attempting to establish eye-movement behavior as an interesting window into audiovisual speech perception (Lansing & McConkie, 1994). Researchers have been concerned more with the final product of speech perception than with the means by which it is achieved. This is arguably a sensible course. For example, perceivers’ eye motion may tell us very little about audiovisual perception if all that is required for visual enhancement is that the salient regions of the face fall within a certain angle of view. That is, the active role of the visuomotor system may be only to point the eyes at a speaker’s face. If so, then we will have to continue to rely on more traditional identification and discrimination tasks for information about audiovisual perception.

Another potential problem with using eye-movement behavior to examine perception is that the points of visual fixation and visual attention need not coincide. For example, subjects can accurately detect characters in the periphery of the visual field while fixating on another target (Jacobs & Lévy-Schoen, 1988; also, see Posner, 1980). Another often remarked example is the ability of deaf signers to decipher hand signs in the visual periphery while foveating on the interlocutor’s face (Swisher, Christie, & Miller, 1989). This suggests multiple foci of attention whose relations with the location of the fovea are quite complex. Finally, nonlinguistic factors may also enhance speech intelligibility. For example, visual orientation may enhance auditory processing, as suggested by the increased intelligibility that occurs when perceivers are allowed to orient to the device (e.g., loudspeaker) conveying the acoustic source (Reisberg, McLean, & Goldfield, 1987).

It has been argued that the phonetically relevant visual information is largely, if not entirely, the by-product of generating the speech acoustics (Vatikiotis-Bateson, Munhall, Hirayama, Lee, & Terzopoulos, 1996). The spatiotemporal behavior of the vocal tract articulators involved in sound production—lips, jaw, and tongue—constrain the shape and time course of visible orofacial behavior. Indeed, the face below the eyes is the visible surface of the vocal tract. Its motion provides visible attributes of speech production that are coherent and coextensive even at a segmental level with the speech acoustics (Vatikiotis-Bateson & Yehia, 1996) and positions of invisible vocal tract articulators such as the tongue (Vatikiotis-Bateson & Yehia, 1997).

A common way to demonstrate the visual enhancement of speech perception has been to manipulate the level of acoustic masking noise. Intelligibility is maintained better at higher levels of masking noise when both visual and acoustic cues are available (Sumby & Pollack, 1954). Analysis of audiovisual stimuli in such studies has typically been restricted to sentence-length or shorter utterances (Demorest & Bernstein, 1992; MacLeod & Summerfield, 1990; Summerfield, 1979; cf. Reisberg et al., 1987). In this study, we chose to examine longer utterances, scripted as conversational monologues, in order to allow longer term patterns in the eye-movement behavior to emerge.

Prior to the study reported here, a pilot study (Vatikiotis-Bateson, Eigsti, & Yano, 1994) was conducted that demonstrated that using longer conversational monologues (presented with and without visual stimuli at different masking noise levels) produced the expected effects on perceptibility. That is, intelligibility of noise-masked speech improved when perceivers could observe the speaker’s moving face. The pilot study also revealed that presentation of a roughly life-size talking head at a 1-m monitor-to-subject distance made it difficult to distinguish when the subject was gazing at the speaker's mouth, nose, or eyes. At that distance and image size, the diameter of the visual fovea (about 1º of arc) was about one third the distance between the eyes and mouth. Thus, the small shifts of gaze required to move between eyes and mouth were quite close to the effective dynamic resolution of the eye-tracking system, approximately 0.5º.

In the present study, therefore, image size and noise level have been manipulated across a range from life size to about five times life size. At the larger image sizes, fixation targets can be reliably distinguished. Indeed, at the largest image size, the angle of view between the speaker's eyes and mouth is about 11º, somewhat beyond the range of visual hyperacuity (Carpenter, 1988; Polyak, 1941). Thus, in addition to providing methodological comparison with life-size images, the larger projected images should induce eye-movement patterns from which the relative importance of fixating on the eyes or the mouth during audiovisual perception can be determined. If perceivers need the mouth and/or eyes to be in sharp focus, then, at larger image sizes, they are more likely to commit to one or the other fixation point. At the very least, the unnaturally large separations between eye and mouth targets induced at larger image sizes should affect either the patterning of eye-movement behavior or the intelligibility results.

Finally, being situated in a Japanese laboratory with access to a fairly large population of expatriate native English speakers, data were collected for both Japanese and English perceivers. This afforded an excellent opportunity to identify common attributes of eye-movement patterning during adverse listening conditions despite quite different linguistic stimuli and perhaps different cultural constraints concerning direct eye contact. There is some evidence that the Japanese sound system does not provide as many visible distinctions in a McGurk task, as do languages such as English (Sekiyama & Tohkura, 1993; Sekiyama, Tohkura, & Umeda, 1996). In addition, it is often remarked anecdotally that the Japanese avoid direct eye contact in certain face-to-face interactions. Thus, in the present study, it was possible that Japanese-speaking perceivers would exhibit tendencies to look less at the speaker's eyes or would choose different facial landmarks for fixation than would their English-speaking counterparts.

METHOD

Audiovisual Stimuli

Two stimulus videotapes were made using a digital recording system (Sony Betacam Model BVP-7), one each for a speaker of standard Japanese (Tokyo) and a speaker of standard American English. Each tape showed the head and shoulders of the speaker as he read a series of 32 conversational monologues.[1] The monologues were 35–45 sec long and were scripted to be plausible within the context of a social gathering, such as a party.

The audio tracks of the monologues were mixed with acoustic masking noise, consisting of multilingual voices and music recorded at a party, to give four levels of masking noise, ranging from no noise to high noise.[2] The pilot study (Vatikiotis-Bateson et al., 1994) showed a qualitative change in eye-movement patterning when masking noise levels were so high that the audiovisual stimuli were unintelligible. For this study therefore, masking noise levels were set so that the audiovisual stimuli at the highest noise level would still be somewhat intelligible. The audio remix and video were transferred to Super VHS format videotape in a pseudo-random order, giving four blocks of eight monologues, where each block contained two monologues at each of the four masking noise levels. Two multiple-choice questions were added to the stimulus tape at the end of each monologue.

Equipment and Recording Procedures

The subjects were seated 1.3 m from a 2 × 3 m back-projection screen. A high-quality liquid crystal projector (Sharp XV-S1Z) was used to present stimulus videos at four image sizes, ranging from approximately life size (scaled to a reference subject–speaker distance of 1 m) to about five times life size. The average vertical angles subtending the stimulus speaker's eyes and mouth were 5.0º, 6.5º, 8.4º, and 10.5º for the four image sizes. Image intensity was not adjusted; therefore, intensity and sharpness of the image decreased as projected size increased.

Horizontal (x) and vertical (y) motions for both eyes were recorded using an infrared, corneal edge-detection system mounted on clear plastic goggles (Takei Co.; see Yamada, 1993). The spatial resolution of the system was about 0.5º (<0.87 cm linear). System calibration consisted of orienting (DC offset) and scaling (gain) the range of detected eye positions relative to a grid of LEDs (50 × 50 cm), under computer control during a set of tracking procedures. Twelve-bit A/D conversion of the audio track and the four channels of eye-position data (vertical and horizontal × 2 eyes) was done at 1000 Hz using a Data Translation board (DT3382) controlled through a VAX-4000 computer.[3] As a quick means of identifying gaze location and checking calibration stability during the course of the experiment and later during data processing, the eye-movement data and the stimulus video were superimposed on a second videotape using a scan converter (Chromatek 9120).

After recording, the eye-position data were numerically filtered at 40 Hz (moving triangular window) and then down-sampled from 1000 to 200 Hz. Given that blinks often occur at the onset of gaze location changes (for review, see Grüsser & Landis, 1991; Leigh & Zee, 1991), they were edited out as smoothly as possible from the horizontal and vertical position data for both eyes.

Design and Procedure

The experiment protocol consisted of presenting a block of eight conversational monologues at each of the four projected image sizes. Since speaking typically causes the goggles to shift position thereby destroying the calibration, the subjects were instructed to try to understand the monologues and to answer the multiple-choice questions using hand gestures.

A calibration trial was recorded at the beginning of each of the four trial blocks and at the end of the fourth block. The subjects were shown a still image of the stimulus speaker's face projected at the appropriate image size. A rectangular frame consisting of two orthogonal sets of parallel lines was superimposed on the still image by a second projector. The subjects were instructed to fixate sequentially on the four intersections of the projected lines and on the speaker's eyes and mouth (see Figure 1). Also, prior to each trial within a block, the subjects traced the four intersections of the projected frame, but not the eyes and mouth. After each trial, the subjects answered the multiple-choice questions projected on the screen. Whenever calibration was lost during the experiment (4 subjects), the system was recalibrated, and the last trial was re-recorded.

Subjects

Five native speakers of English (henceforth referred to as the *English* subjects) and 5 native speakers of Japanese between 23 and

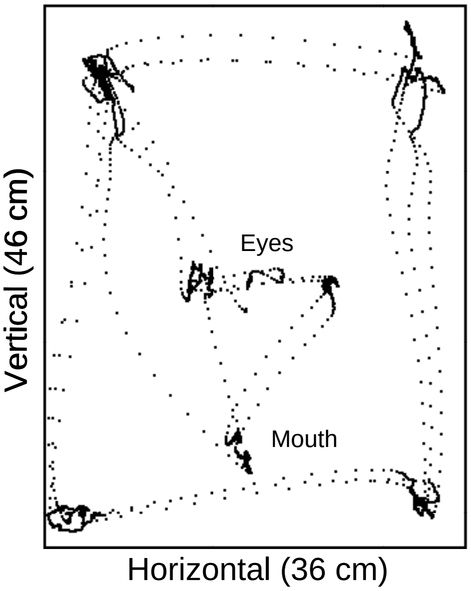

Figure 1. The two-dimensional position of the left eye is plotted for a preblock calibration trial (image size 3). The subject sequentially fixated on the four corners of a projected grid, the eyes, and the mouth.

36 years of age participated in the study. The English subjects were not dialectally uniform: 4 were American, and 1 was British. Although not all Japanese subjects were from the Tokyo dialect area, the preferred Tokyo dialect prevails in schools and the media and, therefore, poses no intelligibility problems. The subjects reported no hearing or speech problems, and all had adequate vision for reading questions from the projection screen. Because of the nature of the eye-tracker, people with contact lenses or particularly large-frame eyeglasses could not be used.

RESULTS

In what follows, the results of four quite different analyses are presented. First, measures of spatial location and variability were used to determine the principle targets of foveation and to assess the effects of masking noise and image size on target choice. Second, the data were analyzed for evidence of sequential patterning in the saccadic shifts among the principal targets of foveation. Third, the eye-motion data for 2 subjects were analyzed with respect to the segmental acoustics for evidence of phoneme identity effects on the eye kinematics. Finally, subject responses to the posttrial multiple-choice questions were used to confirm the settings of the acoustic masking noise.

The analyses of eye motion were all based on the horizontal and vertical position data for one eye. Statistical reliability of the various spatiotemporal measures of eye motion and the intelligibility scores was tested for the two language groups using three-way analyses of variance (ANOVAs), with repeated measures on masking noise level and image size. Measures for the two tokens of each noise level × image size condition were averaged, resulting in 16 cells for analysis. Error bars in the graph figures denote the standard error of the mean (*SEM*).

Where Do Perceivers Gaze During Audiovisual Perception?

In this section, we describe where perceivers gazed during the audiovisual perception task, and how their gaze patterning was affected by noise level and image size. Among other things, we wanted to know the extent to which perceivers need to gaze directly at the mouth in order to extract phonetically relevant visual information. Specifically, we assessed the effects of masking noise level and image size on the relative time the perceivers spent gazing at the mouth versus the eyes, and how often they shifted their gaze to the mouth. Two measures were used to test this: the relative proportion of a trial that the perceiver gazed at the mouth, and the number of saccadic gaze transitions between the eyes and the mouth.

The eye-position data were assigned to the five bins shown in Figure 2. Bins 1 and 2 denote the regions left and right of midline, where the gaze was above the brow line. Bins 3 and 4 are for the left and right eyes, as seen by the perceiver. Bin 5 includes the lower face region around the mouth. The associated calibration trial for each image size and trial-specific corrections were used to determine the vertical midpoint between the stimulus speaker's eyes and mouth for each trial. The vertical line separating bins 1 and 3 from bins 2 and 4 was defined at the horizontal midpoint between the two eyes.

Since less than 1% of the position data fell in bins 1 and 2, all samples falling above the vertical midpoint separating the eye bins from the mouth bin were assigned to a generic eye bin. The proportion of a trial in which the gaze was fixated on the mouth was then calculated for

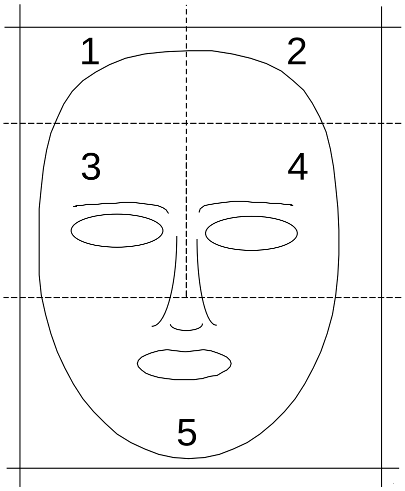

Figure 2. Schematic diagram shows the divisions of the image into five bins. Bins 1 and 2 divide the area between the hairline and the top of the image along the vertical midline of the face. The horizontal boundary between bins 3 and 4 above and bin 5 below was defined at the midpoint between the two eyes and the mouth.

each condition by dividing the number of samples in the mouth bin by the total number of samples in the trial. An ANOVA of this proportion gave a single main effect of noise level [$F(3,24) = 5.87, p < .01$] and no interactions. As plotted in Figure 3, the proportion of the trial that perceivers gazed at the mouth increased with noise level, ranging from about 35% at the *none* and *low* noise levels to 55% at the *high* noise level.

The number of transitions or saccades per trial that occurred between the eyes and mouth was easily computed, since the subjects fixated primarily on either the mouth or the eyes. Trial-length differences were normalized by expressing each trial's number of samples as a fraction of the longest trial. The number of transitions for each trial was then multiplied by the resulting scale factor and analyzed for the effects of noise level and image size. Again, there was a main effect only for noise level [$F(3,24) = 5.19, p < .01$] in which the number of saccades decreased as noise level increased (see Figure 4).

Although duration of gaze fixation was not measured directly in this study, these two results afford an indirect estimate: The duration of gaze fixations on the mouth was about 3.5 times larger at the highest level of acoustic masking noise than in the acoustically clear condition. That is, as the number of transitions denoting fixations within the target regions was halved at the highest noise level, the proportion of "time" spent on the mouth was increased by 60% from .35 to .55 of the trial.

When Do Perceivers Gaze at the Mouth?

In the following subsections, two analyses are presented that were intended to address more temporal aspects of the eye motion. The first describes the patterning of gaze fixation sequences for various sequence lengths. The second analysis concerns the correlation between the location of the perceiver's gaze at a given point in time and the phonetic content of what the speaker was saying.

Patterning of gaze sequences. In this section, the patterning of gaze sequences evoked as the perceivers shifted their gaze among the five facial regions is examined briefly (for details, see Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1996). Since this phenomenon has not previously been described in the literature for an audiovisual perception task, the intent here is to provide a basic description and to address several questions: Do perceivers employ identifiable subsets of the possible gaze sequences? If so, what are they, of what length, and how are they affected by changes in the audiovisual environment?

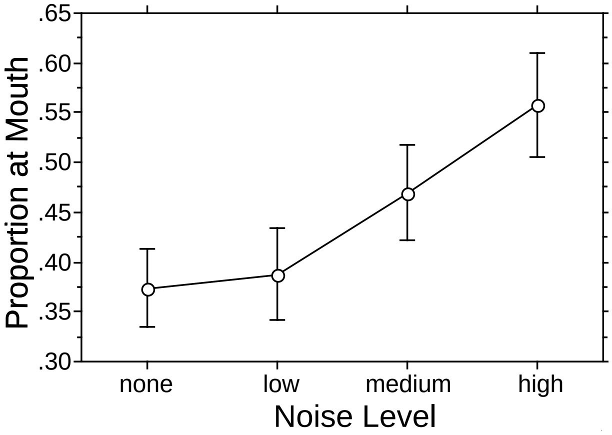

Figure 3. The proportion of fixations during a trial falling in the mouth bin is plotted as a function of masking noise level.

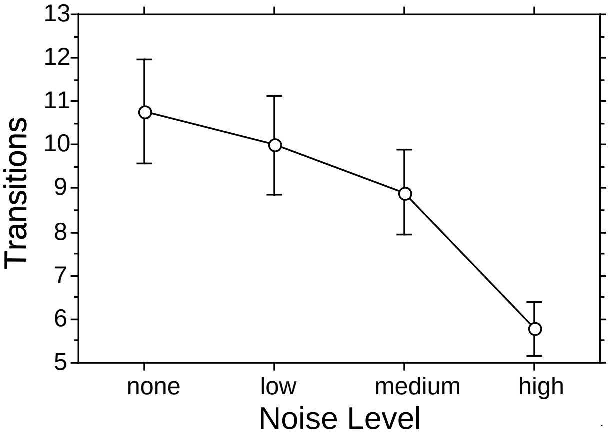

Figure 4. Means for the number of gaze transitions between eyes and mouth within a trial are plotted as a function of noise level.

Initially, the eye-position data were assigned to the five facial bins shown in Figure 2. However, all patterns accounting for 1% or more of the elicited sequences consisted entirely of fixations falling within the three eye and mouth bins. Use of only three bins more accurately reflects the distribution of "high-frequency" pattern types and results in a much smaller set of possible pattern types. As shown by the following relation, the number of possible pattern types (PTs) depends on the sequence length (SL) and the number of analysis bins (B). Using three bins,

$$PT = B * (B - 1)^{(SL - 1)} = 3 * 2^{(2,3,4,5,6)}.$$

The basic results for each language group are given in Table 1 as a function of sequence length; these results were further analyzed using chi-square tests for the effects of noise and image size. The elicited sets of pattern types and frequency counts are given for each language separately and in combination, along with the total number of possible pattern types. The noteworthy finding here is the small number of patterns elicited, even at longer sequence lengths where many more patterns are possible, and the high degree of overlap between the pattern sets used by the Japanese and English subjects. The combined total of observed pattern types reached a maximum of only 33 at sequence length 5, and even less pattern variability was elicited for sequence lengths 6 and 7. Note that the difference in the overall numbers of patterns (count) observed for English and Japanese subjects was due to the longer duration of stimulus monologues for the English conditions.

The most common patterns were repetitive sequences involving transitions between the two eyes (e.g., 3-4-

Table 1
Gaze-Sequence Patterns and Pattern Counts as a Function of Gaze-Sequence Length (SL) and Language

SL	Condition	English	Japanese	Combined	Possible
3	Patterns	12	12	12	12
	Count	3,560	3,078	6,638	
4	Patterns	23	21	24	24
	Count	3,331	2,823	6,154	
5	Patterns	29	26	33	48
	Count	2,718	2,355	5,073	
6	Patterns	27	20	27	96
	Count	1,883	1,699	3,582	
7	Patterns	18	14	20	192
	Count	1,123	1,251	2,374	

Note—Results are presented for analysis of the three eye and mouth bins. Aggregate results for English and Japanese perceivers are given under "Combined."

3-4), followed next by a pattern involving one of the eyes and the mouth (e.g., 3-5-3-5). This ranking of pattern types persisted for all pattern lengths (shown for sequence length 3 in Figure 5). Comparison of means showed no consistent preference for which of the two targets initiated a repetitive sequence (e.g., 3-5-3-5 vs. 5-3-5-3). In most cases, means were nearly identical, and, when they were not, the difference was not predictable. The next most common were patterns consisting of sequences combining a repetitive eye–eye sequence with a transition to the mouth and perhaps back again (e.g., 3-4-3-5). The least common of the high-frequency patterns were those tracing the apices of the eye–mouth triangle in a circular fashion (e.g., 3-4-5-3). English subjects consistently displayed a wider variety of patterns than did the Japanese for all sequence lengths except 3 (see Vatikiotis-Bateson, Eigsti, et al., 1996, for details).

Table 2 shows pattern frequencies for the different sequence lengths as a function of noise level on the left and image size on the right. There were fewer patterns at higher noise levels for subjects of both languages, which agrees with the smaller number and longer duration of gaze fixations at higher noise levels. The differences between lower and higher noise levels were less extreme for the Japanese subjects than for the English subjects: At lower noise levels, fewer patterns were elicited for the Japanese subjects; at higher noise levels, fewer patterns were produced by the English subjects. As can be seen in the right half of Table 2, image size effects on pattern frequency did not vary consistently across the four projected image sizes. Furthermore, the inconsistency differed for the two language groups.

In general, for both language groups, the diversity of patterns reduced as noise level and image size increased. A consequence of this reduction in diversity was the disproportional increase in repetitive eye–mouth gaze shift patterns involving predominantly one eye rather than the other. At higher noise levels, English subjects gazed more at the left eye, whereas the Japanese subjects gazed more at the right eye. The larger image sizes elicited an increase in repetitive eye–mouth gaze shift patterns, particularly the patterns involving the right eye (e.g., 3-5-3-5-3).

Phonetic correlates of gaze location. In attempting to discern a causal link between the audiovisual stimulus and perceiver eye-motion patterning, we tested the possibility that gaze fixations on the mouth might be correlated with the visual salience of the phonetic gestures being produced. That is, bilabials, labiodental and alveolar fricatives, and high vowels (spread /i/ and rounded /u/) have strong visual correlates, which precede the corresponding segmental acoustics by as much as 150 msec (Cathiard, Lallouache, & Abry, 1996). Perceivers could conceivably make use of such phoneme-specific information to enhance perception.

The eye-movement data for 2 subjects (see note 2) were coded for target location and for the identity of the preceding, current, and following phonetic segment. A correlation with a temporally prior segment would suggest probabilistic prediction of visible oral events based presumably on a combination of prior acoustic and visual events, whereas correlations with the following segment would suggest a more simply reactive visual response. A correlation with the current segment could imply either prediction or pretuning of the oculomotor system to reduce reaction time (for review, see Carpenter, 1988) or a combination of predictive and reactive phenomena. How-

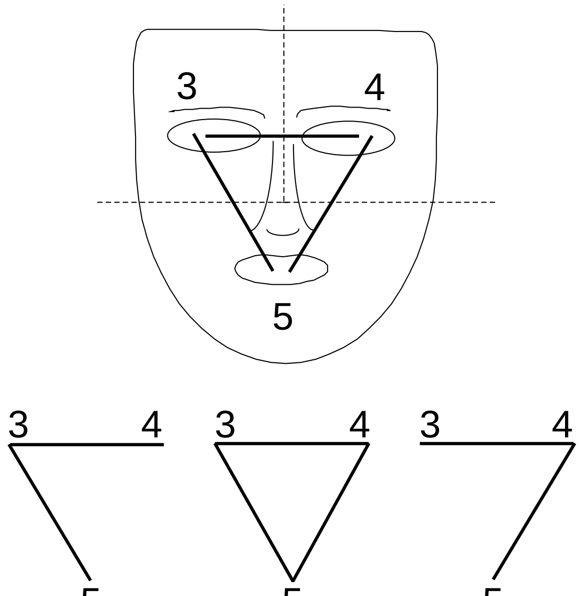

Figure 5. Overlaid on the schematic face are the most common gaze-sequence patterns involving repetitive transitions between just two targets: eye–eye (3-4-3 . . .) or eye–mouth–eye (3-5-3. . ., 4-5-4. . .). The next most common patterns are shown below: at the sides, the pattern templates for repetitive transitions between two targets with an occasional transition to the third target are the most common (e.g., 3-4-3-4-5); in the middle is shown the slightly less common pattern entailing successive clockwise or counterclockwise sweeping of the three targets (e.g., 3-4-5-3-4).

Table 2
Effects of Noise Level and Image Size Condition on Gaze-Sequence Pattern Counts Summarized as a Function of Sequence Length (SL) and Language

		Noise Level				Image Size			
SL	Language	None	Low	Mid	High	1	2	3	4
3	English	1,119	1,039	919	483	891	812	1,109	748
	Japanese	845	894	755	584	603	971	693	811
4	English	1,066	968	859	438	836	766	1,053	676
	Japanese	777	831	689	526	555	903	623	742
5	English	900	804	655	359	708	644	878	488
	Japanese	657	689	571	438	475	751	501	628
6	English	655	582	391	255	504	478	605	296
	Japanese	485	466	435	313	364	527	310	498
7	English	371	332	252	168	344	276	366	137
	Japanese	349	337	337	228	270	377	196	408

ever, no statistically significant correlation between eye position and phonetic identity has been found for any of the comparisons made thus far ($p > .1$).

Characterizing the Details of Eye Motion

In the preceding sections, we examined the proportion of time perceivers gazed at the mouth and eye targets and the patterning of gaze shifts among them. In this section, several aspects of the eye-movement behavior are quantified separately at the eye and mouth targets.

Motion at the eyes. As shown above, the perceivers spent 45%–70% of each trial gazing at the speaker's eyes, depending on the masking noise level. The gaze-sequence results suggest a preference for one eye over the other in repetitive eye–mouth gaze shifts, particularly at higher noise levels and larger image sizes. In this section, this preference is quantified by examining the relative difference in fixation "time" for the two eyes.

Eye-position samples were assigned to bins 3 and 4 for the stimulus subject's right and left eye, respectively (see Figure 2). The values within each bin were summed and normalized for trial-length differences. The proportion of each trial that the perceivers fixated on one eye or the other was calculated as an absolute difference value. Thus, exactly which eye was preferred was not noted. Analysis of the relative difference between eyes gave a main effect of noise level [$F(3,24) = 6.66, p < .01$]. There was also an interaction between noise level and language [$F(3,24) = 4.38, p < .05$], which is shown in Figure 6.

Figure 6 shows that the perceivers gazed predominantly (>70%) at only one of the eyes and that this asymmetrical preference increased by a few percents at higher noise levels. English and Japanese perceivers differed in which of the two higher noise level conditions showed the major increase in the relative difference. For the three noise conditions in which some masking noise was present, the English subjects showed progressively larger relative differences as noise level increased. For the Japanese subjects, the relative difference was highest at the medium noise level.

When coupled with the findings that the total gaze duration on the eyes and the number of transitions decreased substantially at higher noise levels, this result implies that one eye acted as the predominant pivot point for eye–eye and eye–mouth transitions. However, the percentage increase in eye-preference asymmetry was quite small. This suggests that the asymmetry may be fairly independent of changes in gaze duration and saccadic gaze patterning.

Motion at the mouth. As discussed in the preceding sections, masking noise level affected both the duration of fixations on the mouth target and the patterning of the saccadic sequences that the perceivers used to shift their gaze to and from the mouth. In this section, the effects of noise level and image size are assessed more locally by examining the eye-movement behavior only in the vicinity of the mouth. The first measure examined assessed the variability of eye motion. In preliminary assessments of these data (e.g., Vatikiotis-Bateson et al., 1994), it was suggested that noise-level effects on the fine-grained stability of gaze fixation might reflect changes in visual attention independent of increased fixation duration and macroscopic changes in saccadic patterning. The second and third measures examined below suggest that this is probably incorrect. More probable is that variability within the mouth bin is an artifact of changes in the overall movement behavior.

Centroid means were calculated for the data falling within the mouth bin (see Figure 2). The per-trial mouth

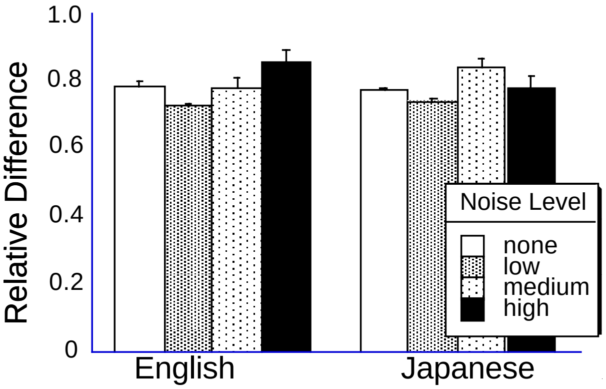

Figure 6. The proportion of fixations on the eyes within a trial attributable to one eye in particular. The relative difference is plotted as a function of noise level and language.

centroid served as the origin for conversion of the sample data from Cartesian (**x**,**y**) to polar (**r**,**q**) coordinates, where **r** denotes the Euclidean distance of a data point from the origin and **q** the angular orientation of the data point relative to the origin. The distance **r** provides a measure of the sample-by-sample deviation from the centroid for the trial. The sum, $\sum$**r**, of all **r** for a given image–noise condition (i.e., across two trials) gives the average deviation from the centroid, **r**, when normalized for differences in trial length.

There was no reliable difference in centroid position for the condition-specific means. However, an ANOVA on the range of motion for the mouth-centered data of 9 subjects (one Japanese subject never fixated on the mouth) revealed main effects of noise level and image size with no interactions. Values of **r** increased at larger image sizes [$F(3,21) = 7.21, p < .01$] and decreased at higher noise levels [$F(3,21) = 5.09, p < .01$]. These two effects are plotted together in Figure 7.

We believe that the increase of **r** at larger image sizes was due simply to the increased size of the mouth target region. The noise level effects, on the other hand, require further investigation to see to what extent they are an artifact of changes in the overall movement patterning. A first step is to compare the contributions of the horizontal and vertical components of the motion with the overall variability. The reduced number of saccades between the eyes and mouth and the longer fixation duration at higher noise levels imply a change in the relative amount of eye-position data associated with the target-to-target motion. Thus, there should be less motion at higher noise levels. Furthermore, since the eye targets are more vertically than horizontally distant from the mouth, changes in the number of shifts between eyes and mouth should affect the vertical component more.

The horizontal component of the motion (**x**) was computed for each sample within the mouth bin by subtracting the horizontal value of the centroid from the raw value of horizontal position. Using the mean Euclidean distance **r**, the ratio (**x**/**r**) was derived denoting the proportional contribution of horizontal motion. An ANOVA yielded no reliable main effects of noise level or image size and no reliable interaction between them. However, a post hoc orthogonal contrast showed a small, but reliable, linear trend in the effect of noise level [$F(1,4) = 4.99, p < .05$]. There was no effect of image size. As shown in Figure 8, the proportion of the horizontal component, which is the larger of the two, decreased by about 6% as noise level increased.

This result is interesting because the overall proportionality of the horizontal and vertical components reflects the geometry of the mouth target—that is, the mouth is about twice as wide as it is high, or 67% and 33%, respectively. However, the weak noise-level effect contradicts the prediction that changes in the amount of eye motion into and out of the mouth target region would have a larger effect on the vertical component than on the horizontal component. This problem is addressed by a third analysis that provides rudimentary evidence that transitions to the mouth from one eye arrive in areas of the horizontally arrayed mouth target different from areas in which transitions from the other eye arrive.

The mouth bin (see Figure 2) was divided into two halves along the vertical midline of the lips.[4] Using the technique described above, absolute differences in the proportion of position data falling on one side of the mouth or the other were computed. Similar to the perceivers' preference to fixate on one eye, a preference was found for one side of the speaker's mouth. An ANOVA showed that the fixation asymmetry increased at higher noise levels [$F(3,24) = 11.82, p < .001$] and that the effect was more pronounced at larger image sizes, as shown by the interaction of noise and image size [$F(3,24) = 2.61, p < .05$]. Furthermore, rather than remain on the same side,

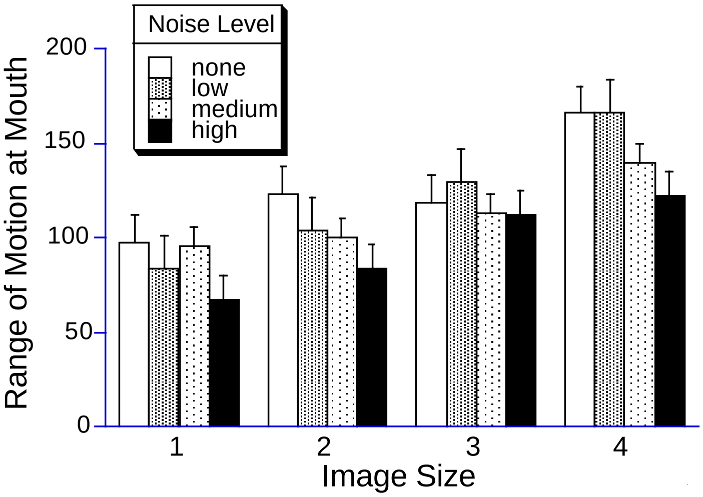

Figure 7. The main effects of noise level and image size on the range of motion around the mouth are plotted together.

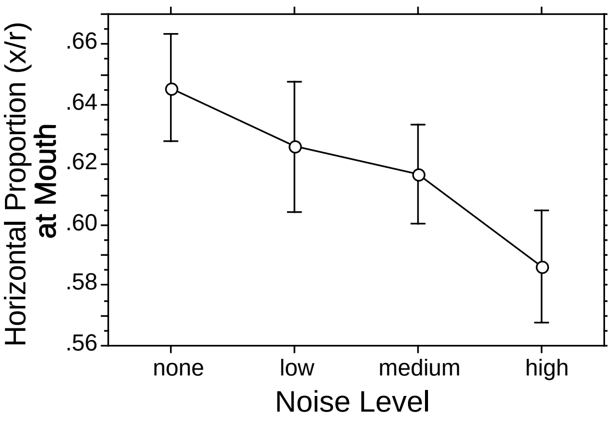

Figure 8. The proportional contribution, x/r, of the horizontal component (x) to the average distance (r) from the centroid is plotted as a function of masking noise level for the mouth region.

eye to the mouth would usually go to the right side of the mouth. Thus, at higher noise levels, the greater tendency to produce one type of eye-to-mouth transition could account for the increased asymmetry shown here as well as the reduced variability shown above.

Stimulus Intelligibility

The main purpose of the stimulus intelligibility test was to ensure, post hoc, that the distribution of masking noise levels was adequate. In mixing the stimulus tapes, we had two goals: (1) to reduce intelligibility evenly across the four masking noise conditions, and (2) to ensure that the high noise level would make perception difficult, but not impossible. The pilot study showed that subjects tended to give up on the task when intelligibility fell below 25%–30% (Vatikiotis-Bateson et al., 1994). Because the more natural-sounding party noise used in this study was not of constant intensity, simple settings of level even to long-term average sound pressure levels could not be trusted. As reported below, the intelligibility results confirmed that the signal/noise ratios were adequately adjusted.

Subject responses to the posttrial questions were used to gauge stimulus intelligibility. These questions consisted entirely of phonetic/lexical identifications, taken from the latter half of each monologue.[5] Questions provided two or three choices, such as "Did the speaker say that the car's interior was *wide* or *white*?" and "Did they stop at the Corn Palace or Bill's Place?" plus "Did not hear" and "Heard, but do not remember." As in the examples given, phonetic distinctions were chosen to be visually ambiguous so as to focus the test on audibility, while no particular care was taken with the lexical contrasts. Two questions were asked after each monologue, and a 2-point scale was used to grade subject answers as either right or wrong.

The effects of masking noise were tested for different stimulus image sizes. An ANOVA showed no effect of language, but there were reliable main effects of noise level [$F(3,24) = 60.92$, $p < .001$] and image size [$F(3,24) = 9.85$, $p < .001$]. Noise level affected intelligibility at each image size; as noise level increased, intelligibility dropped. The interaction between noise level and image size was also reliable [$F(9,72) = 3.10$, $p < .01$] (see Figure 9). As seen in some of the other analyses, the difference between *none* and *low* noise conditions was small, and even reversed for image sizes 2 and 4.[6] The more interesting feature of the interaction is that intelligibility, which was somewhat greater for image sizes 2 and 3, fell off for the *medium* and *high* noise levels at the largest image size, 4. A number of speakers reported after the experiment that image size 2, which was about twice the size of the nearly life-size image size 1, was the "easiest" to watch.

An inherent limitation in the use of conversational monologues as stimuli is the possibility that the monologues themselves are not equally intelligible, which could bias the effects of noise level and image size on intelligibility. An ad hoc perception study designed to address

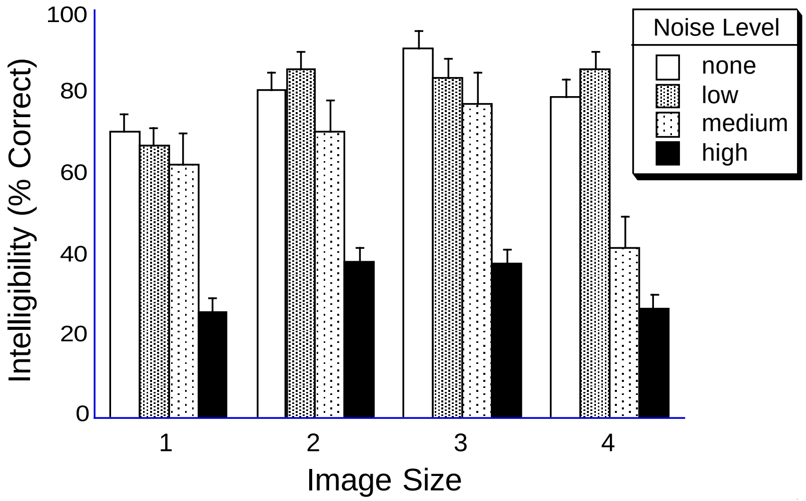

Figure 9. Intelligibility scores (percent correct) are plotted across subjects and language, showing the effects of image size and masking noise level.

this possibility is described in the Appendix. The results show differences in intelligibility for 2–3 of the 32 monologues of each language. However, their random assignment to test conditions in the production study should have canceled out any consistent effects due to the monologues alone.

DISCUSSION

General Remarks

In this study, a number of basic findings have been demonstrated regarding the eye-movement behavior of Japanese and English perceivers during an audiovisual perception task. Most basic with respect to the experimental paradigm is the finding that masking noise level affected perceiver eye-movement behavior at every level of analysis: where perceivers gazed, for how long, and in what spatiotemporal order. Image size, on the other hand, had limited effects on the specific patterning of saccade sequences and on intelligibility scores, but no effects were observed on where subjects gazed or for how long. A methodological benefit of subjects' general resistance to image size effects is that larger-than-life image sizes can be used to enhance the limited resolution of standard infrared eye-tracking systems without substantially altering subject eye-movement behavior. When asked whether or not they found certain image sizes easier to watch, the subjects said they preferred the two middle image sizes over the smallest and largest image sizes. On the basis of this preference, image size 3 has been used in subsequent studies (e.g., Appendix; Eigsti, Munhall, Yano, & Vatikiotis-Bateson, 1995).

The eyes and the mouth were the primary targets of gaze fixation. While not surprising since the video presentation of a talking head puts severe restrictions on the perceivers' field of view, the finding is interesting because the perceivers persistently fixated directly on these targets across a range of audiovisual conditions. At the largest image sizes, for example, the viewing angle between eyes and mouth was about 11º, so the perceivers could not simultaneously view both eye and mouth targets within the region of highest visual acuity (for review, see Carpenter, 1991). Nevertheless, foveation occurred on the eye and mouth targets, rather than somewhere between them, such as the nose.

The subjects spent a larger proportion of time gazing at the eyes than we would have predicted, particularly at the highest noise levels where we expected the perceivers to fixate exclusively on the speaker's lips. At the lowest noise levels, fixations on the eyes accounted for more than 65% of the trial duration. At higher noise levels, gaze fixations shifted more to the mouth, and the frequency of eye-mouth gaze shifts was halved, but fixation on the eyes still occupied 45% of the trial. In part, our expectation may have been conditioned by studies in which subjects were trained speech readers (e.g., Lansing & McConkie, 1994) or were otherwise biased toward speechreading (e.g., the hearing impaired), and the durations of the stimulus trials were relatively short and well marked. Despite the fact that the stimuli used in this study required longer term attention (e.g., to the story line), we also assumed the primary function of interlocutor eye contact to be sociolinguistic, mediating turn-taking, sincerity, and so on. However, even in a noninteractive experimental context such as this, where the sociolinguistic factors attendant to eye contact should be largely irrelevant, their visible correlates (e.g., some eye blinks and eyebrow motions) are strongly coupled to the prosody. Thus, it is probably wrong to think that the perceivers retrieved all of the relevant visual correlates to the prosody and phonetics from the facial deformation patterns associated with jaw and lip motion.

The perceivers showed a distinct preference (70% or more) for one eye over the other, which increased slightly at higher noise levels. Because the relative difference between the two eyes was measured as an absolute value for each trial, we did not report exactly how consistent the subjects were in their choice of preferred eye across trials. However, it is clear from a partial reexamination of the data that perceivers do vary their choice of preferred eye across trials. Also, the preference for a particular eye was demonstrated in the repetitive eye–mouth gaze sequence patterns. This suggests that the preferred eye might act as the pivot for eye–eye and eye–mouth sequences. This issue will be addressed at a later date by examining the trial-specific correlations between preferred eye and preferred eye–mouth transitions.

Finally, the lack of correlation between eye motion and segmental, or syllable-sized, phonetic events was not surprising. Among other things, syllable durations in this study were at the cited lower bound for the time needed by the eye to establish successive fixations, 250–450 msec (Moray, 1993). Fixation durations on the mouth target were usually long, ranging from several seconds to the entire trial. Thus, while there is plenty of evidence from this and previous studies to suggest that perceivers detect phonetically relevant events in the visual stimuli, they do not appear to track or anticipate such events with shifts of gaze fixation.

As with any incursion into a new area of research, hindsight illuminates numerous limitations and peculiarities of the experimental context used. Two of these have been addressed in a subsequent study (Eigsti et al., 1995). First, the linguistic (lexical/phonetic) bias of the posttrial questions could have induced fixation strategies that skewed the effects of acoustic masking noise on eye-movement behavior. That is, the eye-movement behavior associated with attending strictly to linguistic details may be abnormal. Second, the finding reported in the Appendix that the order of monologue presentation has effects on intelligibility suggests that perceiver attention and behavior, particularly with regard to communication, unfold continually through time. Such processes probably cannot be suspended in studies of this sort. In a follow-up to the present study, Eigsti et al. (1995) have shown that presentation order does not mitigate the masking noise effects discussed here. On the other hand, when contrasted with a social judgment task, linguistic discrimination does appear to influence the distribution and patterning of gaze fixations among the eye and mouth targets.

Perceiver Eye Motion and the Production of Audible–Visible Speech

Although this study demonstrated that perceiver eye-movement behavior—a largely voluntary, motor event—can be consistently influenced by manipulating the acoustic and visual parameters of an audiovisual speech perception task, the exact contribution of perceiver eye motion to the process of audiovisual speech perception remains to be discovered. The fact that the perceivers adapted their eye-movement behavior to changes primarily in the acoustic, rather than the visual, environment suggests that the observed effects are not merely epiphenomenal but points up the need to determine what other concurrent visual factors influence eye-motion behavior. For example, the basic patterning of the perceiver eye motion observed in this study is essentially the same as that elicited by simply looking at a face (Yarbus, 1967). The special morphology and functional status of the face in visual perception undoubtedly constrain the role of eye-movement behavior in audiovisual speech perception. Determining how those aspects of the observed behavior that are essential to taking in phonetically relevant visual information are linked to the morphological and other nonphonetic influences on eye motion is beyond the scope of the present study. However, a useful precursor is to examine the what and where of phonetically relevant visual events on a speaker's face. In what follows, a preliminary attempt to do this is made using the results of this study and of related studies aimed at modeling the visible orofacial and audible components of speech production (e.g., Yehia, Rubin, & Vatikiotis-Bateson, in press; for review, see Munhall & Vatikiotis-Bateson, 1998).

The results of this study suggest that fine-grained detection of the perioral structures was not necessary for the visual enhancement effect of the stimulus monologues on perception. This is supported by the failure of the subjects to fixate exclusively on the mouth at the higher noise levels regardless of image size. Gaze fixations on the perioral region would be required for detailed identification (e.g., of lip shape and oral aperture size) if that were where the phonetic information is primarily located.

In some sense, it may be better for perceivers not to foveate continuously on the mouth. In standard descriptions of how the spatially precise fovea and temporally adept periphery coordinate visual detection (e.g., Carpenter, 1988), changes in the visual field are detected peripherally, followed by saccadic shifts of the fovea to, and subsequent fixation on, the point of detection. In this way, new (typically moving) objects in the visual environment are found and identified. Certain extreme phonetic postures might be identified from static images—for example, closed lips (/p/, /b/, /m/), lip rounding (/u/, /o/), or lip spreading (/i/, /s/). However, visual information must be dynamic in order to enhance phonetic perception substantially. For example, Vitkovitch and Barber (1994) have shown that visual enhancement of speech perception begins to deteriorate at frame rates below 16–17 Hz. In the acoustic domain, Remez, Rubin, Berns, Pardo, and Lang (1994) have demonstrated the dynamic nature of speech perception by using sine-wave resynthesis to remove the acoustic complexity of the speech signal while retaining its fine-grained temporal structure.

Thus, by foveating primarily on the eyes during audiovisual perception, spatial acuity might be exchanged for the more accurate temporal detection of perioral events afforded nonfoveally. Since speech is highly overlearned, perceivers probably know quite well the audiovisual information they are seeking. It may be necessary for them only to detect relevant events dynamically and in the right serial order. Furthermore, the acoustic and visual events in speech perception are effectively simultaneous, which may enhance perception by distributing the identification task across the two temporally integrated modalities.[7] As a result, sufficient information about the identity of phonetic events occurring peripherally may be inferred from their timing with respect to other visual and acoustic events and their membership in a closed set of known events. Knowledge of phonotactic, lexical, and other ordering constraints would all contribute to making these events more predictable.

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 speaker's mouth. For the time being, however, we hypothesize that phonetically relevant visual information occurs all over the face, not just the perioral region defined by the lips. This is because the motions of speech articulators such as the lips and jaw, which produce time-varying changes of vocal tract shape (and thereby shape the acoustic output), simultaneously produce dynamic deformations of the entire face.

Recently, the physiology and kinematics of facial motion during production of realistic speech have been examined through analysis of muscle EMG, facial kinematics, and the speech acoustics (Vatikiotis-Bateson & Yehia, 1996; Yehia et al., in press). Three findings are relevant to this discussion. First, the three-dimensional shape and motion of the lips (not necessarily the oral aperture) is correlated at better than 95% with remote regions of facial motion, defined by position markers (or video analysis) on the upper and lower face and the chin. From this, we conclude that different facial regions offer largely redundant motion information. Second, the RMS amplitude of the continuous speech acoustics is equally well recovered from either perioral or more remote facial regions (e.g., 80%) but is better recovered by combining the two regions (e.g., 93%), suggesting that correlates of the segmental acoustics are not distributed uniformly over the entire face. This is somewhat at odds with the usual effort to extract visual phonetic correlates strictly from lip shape and oral aperture size (e.g., Benoît et al., 1992; Montgomery & Jackson, 1983). It also calls into question the common assumption that only a fairly small set of phonemes is visually salient due to the distinctiveness of their labial postures. Third, modeling of the facial motion from the time-varying activity of the orofacial muscles provides kinematic estimations of the remote regions of the face that are as good as or better than estimates of the motion around the lips. This finding points up the importance of understanding the complex anatomy and physiology of the orofacial system.[8]

Thus, for example, motion correlates to lip rounding for the English vowels /u/ and /o/ are visible across the entire face below the eyes. Though individual differences in orofacial anatomy and physiology will ensure slight differences in the actual facial deformation, these are exactly the sorts of differences to which perceivers might rapidly adapt in multimodal interactions. Whatever the actual physical character of phonetically relevant visual information may be, it is not restricted to the perioral aperture. Using the eyes and mouth as fixation points within which potential visual information is redundantly distributed could eliminate the need for a change of foveation strategy when the angular distance between fixation targets is increased. But only up to a point—that intelligibility in high noise decreased at the largest image sizes, when perceivers fixated more on the mouth, could indicate subjects' inability to use the mouth instead of one of the eyes as the primary anchor for their gaze fixation strategies. That is, perceivers may produce habituated eye-movement patterns that serve both phonetic and higher level, sociolinguistic criteria. Because these patterns may be sufficient in a wide range of environments, they may be difficult to change in extraordinary situations, such as our highly artificial perception task.

The Effects of Language on Eye-Movement Patterning

Our expectation that eye-movement behavior would not reflect differences induced by language or presentation style of the stimulus speaker was, on the whole, borne out by the results. Only two analyses showed any language-specific differences in behavior, and both had to do with fixating predominantly on one of the stimulus speaker's eyes. At this point, we have no explanation for this difference; it could have been due to a difference between the two stimulus speakers, such as the amount of head motion or expressive gestures.

The follow-up study of Eigsti et al. (1995) has shown that subjects respond differently to different speakers producing much longer unscripted monologues (e.g., in the number of gaze transitions between the eyes and mouth), but, so far, no differences such as those observed between the two language groups have been found. In particular, no interspeaker effect has been observed that is analogous to the difference in the variety of gaze-sequence patterns used by the Japanese and English perceivers of the present study. This encourages us to attribute the more pronounced tendency of Japanese perceivers to reduce the variety of gaze-sequence patterns sooner as noise level is increased to a combination of cultural and linguistic differences affecting the style and utility of gaze strategies for the two language groups.

Face-to-face communication among Japanese interlocutors provides ample opportunity for audiovisual perception. Only the emphasis on making mutual eye contact appears to be lacking. On the cultural side, the greater variability in gaze-sequence patterns for English perceivers may reflect this more demanding sociolinguistic constraint on mutual eye contact among interlocutors (without actually staring), something independent of strategies that enhance audiovisual perception.

On the linguistic side, Japanese and English speakers may impart different degrees of linguistically relevant visual information, as suggested by Sekiyama and colleagues using the McGurk effect (Sekiyama & Tohkura, 1993; Sekiyama et al., 1996). These investigators suggest that the Japanese phonetic system provides fewer salient visual correlates than does English, on the basis of the weaker tendencies for Japanese perceivers to experience the audiovisual "fusion illusion" when presented with Japanese stimuli than with English stimuli (cf. Massaro, Tsuzaki, Cohen, Gesi, & Heridia, 1993). An important implication of Sekiyama's studies for the results of the present study is that audiovisual processing and its associated mechanisms are structured the same across cultural and linguistic variations. Differences are minor, and behavioral patterns may be easily altered by changing the stimulus, as in Sekiyama's studies.

Summary

In this study, the eye-movement behavior of perceivers during audiovisual speech perception was examined under variable visual and acoustic conditions. The finding that perceiver eye motion was particularly sensitive in a variety of ways to changes in the acoustic environment indicates an active role of the perceiver's motor system in the process of audiovisual perception. Although a variety of analyses were presented whose results apparently assessed the fine-grained structure of eye motion, we concluded that noise level effects were primarily macroscopic, altering the distribution of gazes among the eye and mouth targets. Indeed, only the most macroscopic of analyses—that of gaze-sequence patterns—revealed any difference between the two language groups. From the tendency of the perceivers to watch the speaker's eyes a good portion of the time, even under poor acoustic conditions, we speculated that much of the visual task during audiovisual perception may entail detecting the occurrence of well-learned, phonetically correlated events. Because these events are well known, we further hypothesized that detection can be achieved away from the fovea and that the task is made easier by the dynamic distribution of phonetic information over the entire face. The manner of that distribution was argued to be time-locked to the acoustics and causally linked to speech motor control in that the motion of speech articulators, such as the lips, jaw, and even tongue, simultaneously configures vocal tract and visible orofacial structures. Although many more questions are raised than answered, the present study has set out a methodological and conceptual framework for pursuing a better understanding of audiovisual perception and its relation with speech production.

REFERENCES

Abry, C., Lallouache, M.-T., & Cathiard, M.-A. (1996). How can coarticulation models account for speech sensitivity to audio-visual desynchronization? In D. Stork & M. Hennecke (Eds.), *Speechreading by humans and machines* (NATO-ASI Series F: Computer and Systems Sciences, Vol. 150, pp. 247-256). Berlin: Springer-Verlag.

Benoît, C., Lallouache, [M.-] T., Mohamadi, T., & Abry, C. (1992). A set of French visemes for visual speech synthesis. In G. Bailly, C. Benoît, & T. R. Sawalis (Eds.), *Talking machines: Theories, models, and designs* (pp. 335-348). Amsterdam: Elsevier.

Bertelson, P., & Radeau, M. (1976). Ventriloquism, sensory interaction, and response bias: Remarks on the paper by Choe, Welch, Gilford, and Juola. *Perception & Psychophysics*, **19**, 531-535.

Brooke, N. M., & Summerfield, A. Q. (1983). Analysis, synthesis, and perception of visible articulatory movements. *Journal of Phonetics*, **11**, 63-76.

Carpenter, R. H. S. (1988). *Movements of the eyes* (2nd rev. ed.). London: Pion.

Carpenter, R. H. S. (Ed.) (1991). *Eye movements*. London: Macmillan.

Cathiard, M.-A., Lallouache, M.-T., & Abry, C. (1996). Does movement on the lips mean movement in the mind? In D. Stork & M. Hennecke (Eds.), *Speechreading by humans and machines* (NATO-ASI Series F: Computer and Systems Sciences, Vol. 150, pp. 211-219). Berlin: Springer-Verlag.

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

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

Eigsti, I. M., Munhall, K. G., Yano, S., & Vatikiotis-Bateson, E. (1995). Effects of listener expectation on eye movement behavior during audiovisual perception. *Journal of the Acoustical Society of America*, **97**, 3286.

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

Gray, H. (1977). *Gray's anatomy*. New York: Crown.

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. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology: Human Perception & Performance*, **17**, 278-288.

Grüsser, O.-J., & Landis, T. (Eds.) (1991). *Visual agnosias and other disturbances of visual perception and cognition* (Vision and Visual Dysfunction, Vol. 12). London: Macmillan.

Jacobs, A., & Lévy-Schoen, A. (1988). Breaking down saccade latency into central and peripheral processing times in a visual dual-task. In G. Lüer, U. Lass, & J. Shallo-Hoffman (Eds.), *Eye movement research: Physiological and psychological aspects* (pp. 267-285). Lewiston, NY: Hogrefe.

Jeffers, J., & Barley, M. (1971). *Speechreading (lipreading)*. Springfield, IL: C. C. Thomas.

Lansing, C. R., & McConkie, G. (1994). A new method for speechreading research: Tracking observer's eye movements. *Journal of the Academy of Rehabilative Audiology*, **27**, 25-43.

Leigh, R. J., & Zee, D. S. (1991). Oculomotor disorders. In R. H. S. Carpenter (Ed.), *Eye movements* (Vision and Visual Dysfunction, Vol. 8, pp. 297-319). London: Macmillan.

Luettin, J., Thacker, N. A., & Beet, S. W. (1996). Active shape models for visual speech feature extraction. In D. Stork & M. Hennecke (Eds.), *Speechreading by humans and machines* (NATO-ASI Series F: Computer and Systems Sciences, Vol. 150, pp. 383-390). Berlin: Springer-Verlag.

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

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

Massaro, D. W., Tsuzaki, M., Cohen, M. M., Gesi, A., & Heridia, R. (1993). Bimodal speech perception: An examination across languages. *Journal of Phonetics*, **21**, 445-478.

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

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

Moray, N. (1993). Designing for attention. In A. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness, and control. A tribute to Donald Broadbent* (pp. 53-72). Oxford: Oxford University Press, Clarendon Press.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, **58**, 351-362.

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. Advances in the psychology of speechreading and auditory-visual speech* (pp. 123-139). Sussex: Taylor & Francis, Psychology Press.

Petajan, E. D. (1985). Automatic lipreading to enhance speech recognition. In *Proceedings: Computer Vision and Pattern Recognition* (pp. 40-47). San Francisco: IEEE Computer Society Press.

Polyak, S. L. (1941). *The retina*. Chicago: University of Chicago Press.

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

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lipreading* (pp. 97-114). Hillsdale, NJ: Erlbaum.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review*, **101**, 129-156.

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.

Sekiyama, K., & Tohkura, Y. [I.] (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, **21**, 427-444.

Sekiyama, K., Tohkura, Y. I., & Umeda, M. (1996). A few factors which affect the degree of incorporating lip-read information into speech perception. In H. T. Bunnell & W. Idsardi (Eds.), *Proceedings: ICSLP 96* (Vol. 3, pp. 1481-1484). Newcastle, DE: Citation Delaware.

Smeele, P. M. T. (1996). Psychology of human speechreading. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by humans and machines* (NATO-ASI Series F: Computer and Systems Sciences, Vol. 150, pp. 3-17). Berlin: 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. (1979). Use of visual information for phonetic perception. *Phonetics*, **36**, 314-331.

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.

Swisher, M. V., Christie, K., & Miller, S. L. (1989). The reception of signs in peripheral vision by deaf persons. *Sign Language Studies*, **63**, 99-125.

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

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., & Munhall, K. G. (1996). *Perceiver eye motion during audiovisual perception* (ATR Tech. Rep. No. TR-203, pp. 1-33). Kyoto: ATR Human Information Processing Research Laboratories.

Vatikiotis-Bateson, E., Munhall, K. G., Hirayama, M., Lee, Y. C., & Terzopoulos, D. (1996). The dynamics of audiovisual behavior in speech. In D. Stork & M. Hennecke (Eds.), *Speechreading by humans and machines* (NATO-ASI Series, Series F, Computers and Systems Sciences, Vol. 150, pp. 221-232). Berlin: Springer-Verlag.

Vatikiotis-Bateson, E., & Yehia, H. C. (1996). Physiological modeling of facial motion during speech. *Transactions of the Technical Committee on Psychological and Physiological Acoustics*, **H-96**(65), 1-8.

Vatikiotis-Bateson, E., & Yehia, H. C. (1997). Unified model of audible-visible speech production. In *EuroSpeech '97: 5th European conference on speech communication and technology*.

Vitkovitch, M., & Barber, P. (1994). Effects of video frame rate on subjects' ability to shadow one of two cometing verbal passages. *Journal of Speech & Hearing Research*, **37**, 1204-1210.

Wolff, G. J., Prasad, K. V., Stork, D. G., & Hennecke, M. (1994). Lipreading by neural networks: Visual preprocessing, learning and sensory integration. In J. D. Cowan, G. Tesauro, & J. Alspector (Eds.), *Advances in neural information processing systems 6* (pp. 1027-1034). San Francisco: Morgan Kaufmann.

Yamada, M. (1993). [Analysis of human visual information processing mechanisms using eye movement] (Tech. Rep. No.17). Tokyo: Institute of Television Engineers.

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

Yehia, H. C. Rubin, P. E., & Vatikiotis-Bateson, E. (in press). Quantitative association of acoustic, facial, and vocal-tract shapes. *Speech Communication*.

NOTES

1. The text was printed on cue cards, located just to the side of the camera lens. In order to minimize the effect of the speaker not looking directly into the lens, a medium telephoto was used at a camera-to-subject distance of approximately 2.5 m. In order to increase the naturalness of the monologue presentation, speakers in a subsequent study (Eigsti, Munhall, Yano, & Vatikiotis-Bateson, 1995) produced extemporaneous monologues while looking directly into the lens. Although impressionistically more natural, no discernible differences between the two styles of delivery were found in the analyses.

2. It is unlikely that the party-noise sound track introduced any linguistic bias. Three languages—French, Japanese, and English—were spoken simultaneously by approximately 15 native and non-native speakers of these languages. In addition to the voices, dance music could occasionally be discerned in the background. No individual voices or clear language identifications could be made, and the speech was judged unintelligible by numerous listeners, as well as by the speakers themselves.

3. Due to an oversight, masking noise and speech were mixed on both audio tracks for 8 subjects. Only for the last 2 subjects, where the two tracks were separated and mixed only for loudspeaker presentation, could the clear speech channel be digitized simultaneously with eye movement.

4. There is a potential problem here, which was analyzed further. The eyes form two distinct targets set far enough apart to prevent masking by undetected drifts in calibration, but the mouth is a single target straddling the midsagittally defined boundary between the two bins. An apparent asymmetry in fixation patterns could arise even when subjects fixate on the midline of the speaker's mouth if there is a small error in system alignment. This possibility was ruled out by coding average fixation position for a trial relative to the vertical midline using a 7-point scale: $-1, -.5, -.1, 0, .1, .5, 1$. Average fixations falling on or slightly to one side of midline were coded as ±.1; fixations on the corners of the mouth were assigned ± .5; and those not on the mouth at all were coded as ±1. Symmetrical bimodal distributions were assigned 0, no matter how far off midline individual clusters of data were. Similar to the relative difference results shown in Figure 6, there was an interaction of noise level and language [$F(3,24) = 3.12, p < .05$] for coded distance from the midline. Japanese speakers fixated more toward the left corner of the mouth as noise level increased. English speakers moved toward the right, but the trend was not consistent for the highest noise level.

5. This strategy was chosen in order to reduce the number of "Heard, but cannot remember" and "Did not hear" responses obtained in the pilot study. Although this strategy appears to have been effective, its use raises the possibility that, sometime during the experiment, the subjects became aware that they need not attend to the first half of the stimulus monologue in order to answer the questions. Two ANOVAs with repeated measures were used to test whether or not the distribution of posttrial questions affect perceiver eye-movement behavior. In one, the data for each trial were divided in half; in the other, the 32 trials were divided into first and second sets of 16 trials each. The two halves of each new data set were then compared for the proportion of time spent gazing at the mouth. Although the main effect of noise level was reliable in both reanalyses ($p < .01$), subdivision of the data had no effect or interactions ($p \approx .1$).

6. In our experience, these reversals rarely reach statistical reliability, but they occur quite often as trends in the means. A plausible, but untested, account is that the presence of low-level masking noise helps focus subjects' attention on the audiovisual task without substantially interfering with intelligibility of the stimulus.

7. Actually, in the production of obstruents involving the lips, such as /p/, /v/, /θ/, /z/, there are clear visual correlates that may precede the acoustics by as much as 150–200 msec (for discussion, see Abry, Lallouache, & Cathiard, 1996). In such cases, detection and identification may be primarily visual, since the acoustics are either delayed and/or difficult to identify.

8. The orofacial musculature is a maze of highly interdigitated and usually small fiber bundles (Gray, 1977). For example, the muscles surrounding the upper and lower lips—orbicularis oris superior (OOS) and inferior (OOI), respectively—have no skeletal attachment. Instead, they act as a floating anchor to at least a dozen other muscles that radiate outward and are associated, for example, with smiling (risorius), upper lip raising (levator labii superior), lower lip lowering (depressor labii inferior), and protrusion (mentalis). Contraction of OOS and OOI, which brings the lips together, also exerts pull on all the muscles attached to them. The action of one muscle almost invariably impinges on other muscles, thus distributing the effects of their actions over a wider range than would be expected from consideration of their independent structure (e.g., length, orientation, and primary skeletal attachments). The effects of muscle action on the posture and motion of facial landmarks are further diffused once the damped connective fascia and relatively stiff outer skin layers are considered.

APPENDIX

A concern that arises with the use of naturalistic monologue stimuli is that inherent differences in length, phonetic and syntactic structure, and lexical–semantic content may affect intelligibility independently of the experimentally manipulated noise level and image size conditions. Therefore, a perception study was conducted in which 40 subjects, 20 Japanese and 20 English, were paid to participate. Perceivers of each language were shown the same monologues as used in the production study, but all 32 monologues were presented with only one masking noise level (medium) and at one of the larger image sizes (3). Perceivers for each language were divided into four groups of 5. Monologues were presented to each group in a different order (the original and three new orders). Subjects were tested individually or in small groups of 3.

Table A1
Mean Monologue Intelligibility Scores (0–1) and
Standard Errors of the Mean for Japanese and English Perceivers

	English								Japanese							
	Size 1		Size 2		Size 3		Size 4		Size 1		Size 2		Size 3		Size 4	
Noise Level	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*
None	.43	.05	.50	.04	.52	.06	.51	.06	.42	.06	.62	.04	.66	.04	.62	.04
Low	.49	.04	.52	.04	.41	.06	.48	.05	.51	.05	.68	.04	.67	.04	.69	.04
Mid	.48	.05	.43	.04	.45	.05	.44	.04	.62	.04	.67	.04	.83	.03	.44	.04
High	.25	.04	.57	.04	.64	.06	.43	.05	.66	.05	.72	.04	.65	.04	.63	.03

Note—Although the results were elicited at one image size and one masking noise level, they are tabulated here to reflect the assignment of the monologues to noise level and image size conditions in the production study.

The effects of monologue and presentation order on intelligibility were analyzed using an ANOVA with repeated measures. For each presentation order, responses were pooled for the two monologues of each production-study condition (16 cells). There were main effects of monologue [English, $F(15,240) = 5.37, p < .0001$; Japanese, $F(15,240) = 10.44, p < .0001$]. For each language, mean intelligibility scores for two or three of the monologue pairs were markedly different from the mean.

In order to see how the difference among monologues might have confounded the results of the production study, Table A1 is organized to show the mean scores for the perception study as a function of the noise level and image size conditions used in the production study. For both English and Japanese, two deviant pairs fell on opposite sides of the mean intelligibility score. Since these monologue pairs were associated with the same noise level conditions (but different image sizes) in the production study, any inherent differences in monologue intelligibility would not interact with the effects of noise level on intelligibility. Also, a third monologue in the Japanese series had a much lower than average intelligibility score. Since this case corresponded to a no-noise (none) condition in the production study, its poor intelligibility would not be a problem, because it would lead to underestimation of the difference between noise level conditions.

Both language groups showed an interaction of monologue and presentation order [English, $F(45,240) = 2.74, p < .0001$; Japanese, $F(45,240) = 4.40, p < .0001$]. For Japanese, the main effect of order was not reliable ($p > .05$), and the interaction was due to one presentation order giving intelligibility scores markedly lower than the other three. For English, the main effect of order was reliable [$F(3,16) = 5.19, p < .02$], with mean intelligibility of the four presentation orders distributed evenly between 35% and 60%.

This last result is interesting because it suggests that relatively long term differences (on a scale of minutes and hours) in event sequences affect perceiver performance in word identification tasks. Nevertheless, it must be remembered that the intelligibility scores in both this and the production study were based on identification of only a few words contained within 100- to 130-word monologues, whose masking noise was not uniform. Even so, the similarity of intelligibility scores for the two language groups of the production study (Figure 9), as well as the small perceptual variability among monologues when tested at a consistent noise level, suggests that masking noise level was the primary determinant of intelligibility in the production study.

(Manuscript received November 26, 1996;
revision accepted for publication July 16, 1997.)