TR-H-203

Perceiver Eye Motion during Audiovisual Speech Perception.

Eric VATIKIOTIS-BATESON, Inge-Marie EIGSTI
(Rochester Univ.), Sumio YANO (NHK) and
Kevin MUNHALL (Queen's Univ.)

1996.11.13

ATR人間情報通信研究所
〒619-02　京都府相楽群精華町光台2-2　TEL: 0774-95-1011

ATR Human Information Processing Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
Telephone: +81-774-95-1011
Fax : +81-774-95-1008

PERCEIVER EYE MOTION DURING AUDIOVISUAL SPEECH PERCEPTION

Eric Vatikiotis-Bateson
ATR Human Information Processing Research Laboratories, Kyoto, Japan

Inge-Marie Eigsti
University of Rochester, New York

Sumio Yano
NHK Research Laboratories, Kinuta, Japan

Kevin Munhall
Queen's University, Kingston, Canada

Abstract

The eye movements of subjects 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. In the anlaysis of gaze fixation sequences, e.g., left eye to mouth to left eye to right eye, English perceivers displayed more variety of gaze sequence patterns 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.

Perceiver eye motion During Audiovisual Speech Perception

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" (Jeffers & Barley, 1971; Gailey, 1987), 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; 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 entails the perceiver integrating spatially disparate acoustic and visual events, 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 primarily on dynamic rather than static characteristics of facial images. For example, Vitkovich & 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., Johnson, Rosenblum, & Saldaña, 1994; 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; Wolff, Prasad, Stork & Hennecke, 1994). A second underlying assumption of all branches of speechreading research has been that for both human and machine viewers the sought-after visual parameters should be measured with as much precision as possible (Benoît et al., 1992). 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 towards which they may not foviate 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 audiovisual behavior of speakers. 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, we have 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 is to address these questions from the perceivers 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 (Lansing & McConkie, 1995). Researchers have been concerned more with the final product of 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). This suggests multiple foci of attention whose relations with the location of the fovea are quite complex. Finally, non-linguistic 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 (Reisburg, 1987).

We have 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, 1996a). In addition, to the extent that we have been able to observe them (Hirayama, Vatiktios-Bateson, Gracco, & Kawato, 1994; Vatikiotis-Bateson & Yehia, 1996b), the underlying neuromuscular constraints on orofacial motion and speech motor control are similar to those observed across the full spectrum of biological movement systems, including the eyes and limbs (Zangemeister & Stark, 1981; Poizner, Bellugi, & Klima, 1990; Kelso, Vatikiotis-Bateson, Saltzman, & Kay, 1985)

It is possible then that voluntary eye movements and orofacial motions during speech may provide a common neuromotor substrate upon which audiovisual speech perception occurs. The purpose of this study is to show that perceiver eye motion is a fundamental, *motoric* component of audiovisual speech perception, and further that it can reveal something useful about the linguistically-relevant events which comprise the speaker's orofacial motion.

The Present Study

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 sentential utterances or shorter (Summerfield, 1979; MacLeod & Summerfield, 1987; Demorset & Bernstein, 1992). 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 run which 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-sized talking head at a one meter monitor-to-subject distance made it difficult to distinguish when the subject was gazing at speaker's mouth, nose, or eyes. For, at that distance and image size, the diameter of the visual fovea (about 1 degree of arc) was only 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 0.5 degrees.

In the current study therefore, image size has been manipulated as well as noise level 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 speakers' eyes and mouth is about 11 degrees, somewhat beyond the range of the hyperacute perifovea — 4.2-9.5 degrees (Polyak, 1941; Carpenter, 1988). Thus, in addition to providing methodological comparison with life-size images, the larger projected images might 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, by recording data for Japanese and English perceivers, the effects on eye movement patterning of quite different linguistic stimuli and perhaps different cultural constraints concerning direct eye contact can be assessed. Although it is not clear exactly how such cultural differences should be characterized or separated from linguistic constraints (Sekiyama & Tohkura, 1993), they may be deeply enough ingrained to affect perceiver responses to McGurk effect stimuli. Thus, even in the highly artificial one-way viewing of this experimental context, Japanese perceivers may exhibit habituated tendencies to look less at the speaker's eyes or choose different facial landmarks for fixation than their English speaking counterparts.

Methods

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 seconds 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 4 levels of masking noise, ranging between no noise and high noise. 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 re-mix and video were transferred to Super VHS format videotape in a pseudo-random order, giving 4 blocks of 8 monologues where each block contained 2 monologues at each of the 4 masking noise levels. Multiple choice questions were added to the stimulus tape at the end of each monologue. The choices consisted of phonetically contrastive words and phrases as well as "heard but cannot remember" and "did not hear".

Equipment And Recording Procedures

Subjects were seated 1.3 meters from a 2 by 3 meter back-projection screen. A high quality liquid crystal projector (Sharp XV-S1Z) was used to present stimulus videos at 4 image sizes ranging from approximately life-size (scaled to a reference subject-speaker distance of 1 meter) to about 5 times life-size. The average vertical angles subtending the

stimulus speakers' eyes and mouth were 5.0, 6.5, 8.4, 10.5 degrees for the 4 image sizes (The actual angles between the center of each eye and the mouth center were slightly larger). Image intensity was not adjusted; therefore, intensity as well as sharpness of the image decreased as projected size increased.

Horizontal (x) and vertical (y) motion 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 0.5 degree (<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 x 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 X 2 eyes) was done at 1000 Hz using a Data Translation board (DT3382) controlled through a VAX 4000 computer.[2] As a quick means of identifying gaze location and checking calibration stability, 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 smoothed at 40 Hz (moving triangular window) and down-sampled from 1000 to 200 Hz. Given that blinks often occur at the onset of gaze location changes (for review, see Gruesser & Landis, 1991; Leigh & Zee, 1991), they were edited out as smoothly as possible from the horizontal and vertical position data for both eyes.

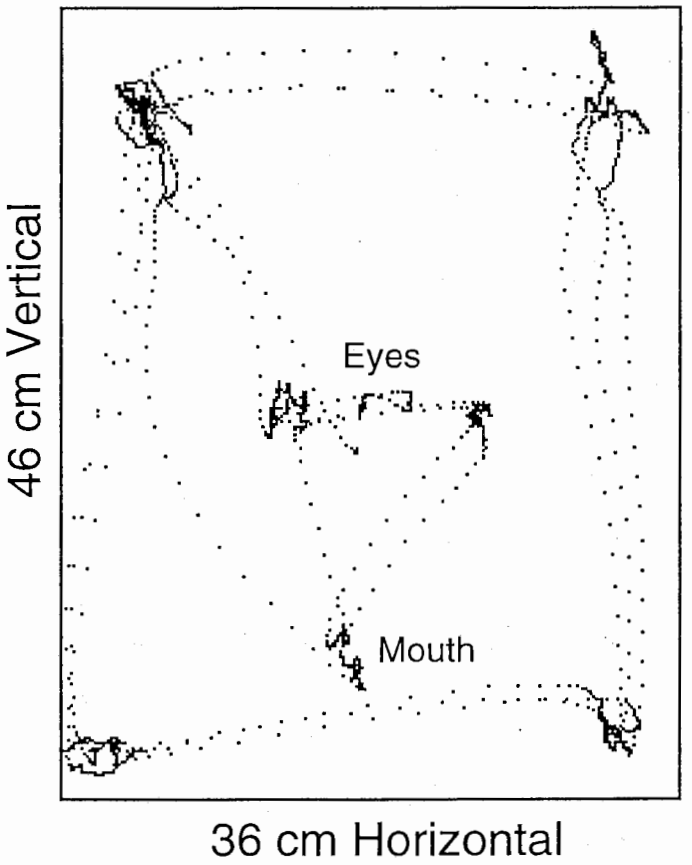

Figure 1. The two-dimensional position of the left eye is plotted for a pre-block calibration trial (image size *3*). The subject sequentially fixated on the four corners of a projected grid, the eyes, and the mouth.

Experiment Design and Procedures

The experiment protocol consisted of presenting a block of 8 conversational monologues at each of the 4 projected image sizes. Since speaking typically causes the goggles to shift position thereby destroying the calibration, 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 4 trial blocks. 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. Subjects were instructed to fixate sequentially on the four intersections of the projected lines and the speaker's eyes and mouth (see Figure 1). Also, prior to each trial within a block, subjects traced the four intersections of the projected frame, but not the eyes and mouth. After each trial, subjects answered the multiple choice questions projected on the screen.

Subjects

Ten native speakers of English (5) and Japanese (5) participated in the study. The English subjects were not dialectally uniform: 4 American and 1 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. 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 eye-glasses could not be used.

Results

In what follows the results of four quite different analyses are presented. First, subject responses to the post-trial multiple-choice questions were used to gauge the phonetic intelligibility of the stimuli. The pilot study had shown that subjects tend to give up on the task when intelligibility is too low (Vatikiotis-Bateson et al., 1994). Therefore, a main concern was to achieve a minimum of 25-30 percent intelligibility. Also, although this was a study of perceiver eye movement behavior, even the limited assessment of intelligibility carried out here provides a common reference with other studies of audiovisual perception. Second, 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. Third, the data were analyzed for evidence of sequential patterning in the saccadic shifts among the principal targets of foveation. Finally, the eye motion data for two subjects were analyzed with respect to the segmental acoustics for evidence of phoneme identity effects on the eye kinematics.

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

Stimulus Intelligibility

Subject responses to the post-trial questions were used to gauge stimulus intelligibility. These questions consisted entirely of phonetic/lexical identifications, taken from the latter half of each monologue in order to minimize memory effects. Questions provided two or three phonetic choices such as "Did the speaker say that the car's interior was *wide* or *white*?", plus "did not hear" and "heard but do not remember". Two questions were asked after each monologue and a two point scale was used to grade subject answers as either right or wrong .

Stimulus intelligibility provided a practical means for checking, *post hoc,* the distribution of masking noise levels. In mixing the stimulus tapes, we had two goals: to reduce intelligibility evenly across the four masking noise conditions, and to ensure that the *high* noise level would make perception difficult, but not impossible. 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 (SPL) could not be

trusted. As reported below, the intelligibility results confirmed that the S/N ratios were adequately adjusted.

The effects of masking noise were tested for different stimulus image sizes. 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 ($F_{[9,72]} = 3.10$; $p < .01$) and is graphed in Figure 2. The difference between *none* and *low* noise conditions is typically small and even reversed for image sizes *2* and *4*. The more interesting feature of the interaction is that intelligibility, which was somewhat greater for image sizes *2* and *3*, falls 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-sized image size *1*, was the 'easiest' to watch.

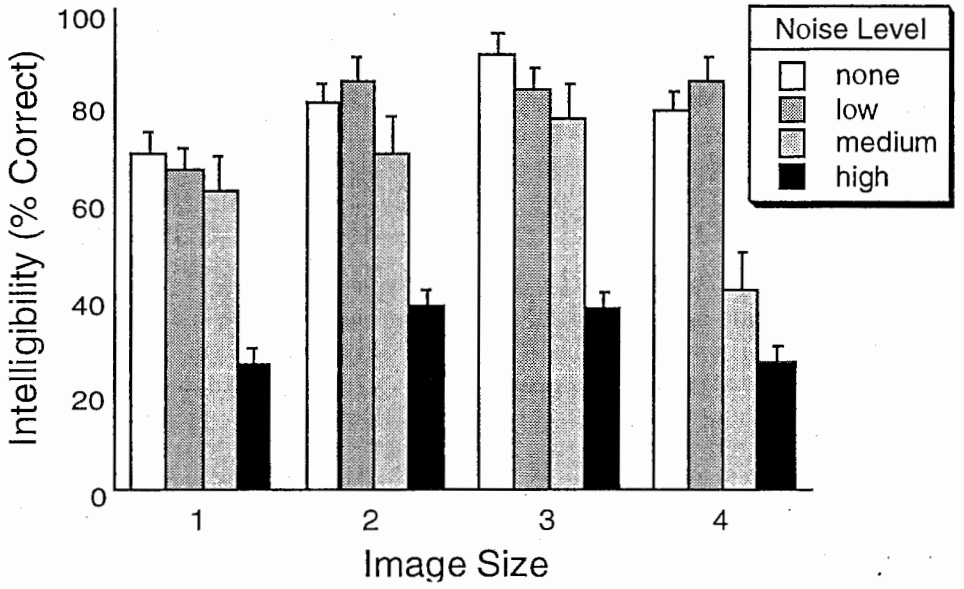

Figure 2. Intelligibility scores (percent correct) are plotted across subjects and language and show the effects of image size and masking noise level.

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. Appendix A describes an *ad hoc* perception study designed to address this possibility. The results show differences in intelligibility for 2-3 of the 32 monologues of each language. However, their distribution over the experimental conditions in the production study should have cancelled out any consistent effects inherent to the monologues themselves.

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 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 3. 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; and 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.

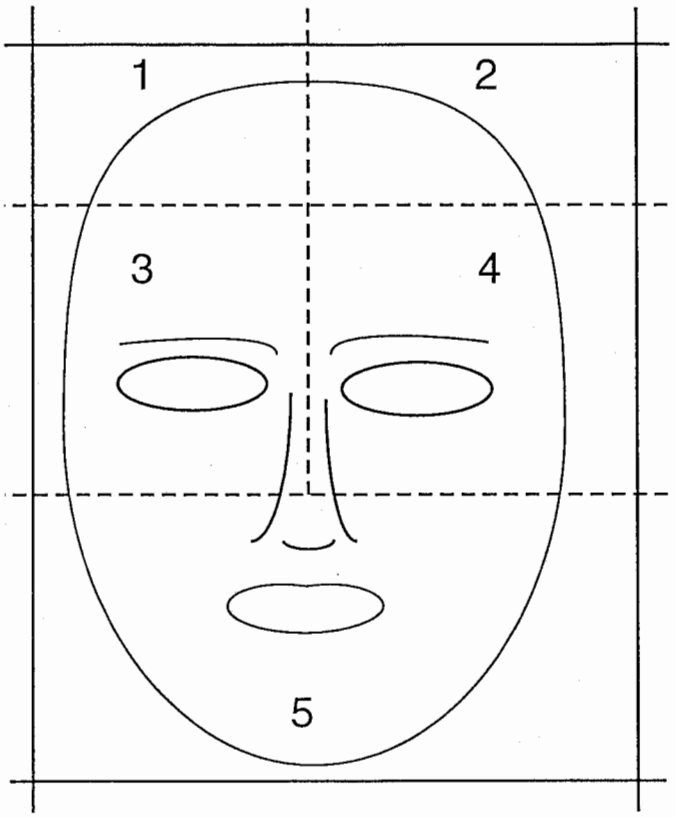

Figure 3. Schematic diagram shows the divisions of the image into five bins. Bins *1* and *2* divide the area between the hairline and the top of the image along the vertical midline of the face. The horizontal boundary between bins *3* and *4* above and bin *5* below was defined at the midpoint between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

Since less than 1 percent of the position 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 each condition by dividing the number of samples in the *mouth* bin by the total number of samples in the trial. 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 4, the proportion of the trial that perceivers gazed at the mouth increased with noise level. The proportion ranged between about 35 percent at the *none* and *low* noise levels and 55 percent at the *high* noise level.

The number of transitions or saccades per trial that occurred between the eyes and mouth was easily computed since subjects fixated primarily on either the mouth or 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 (Figure 5).

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 percent from .35 to .55 of the trial.

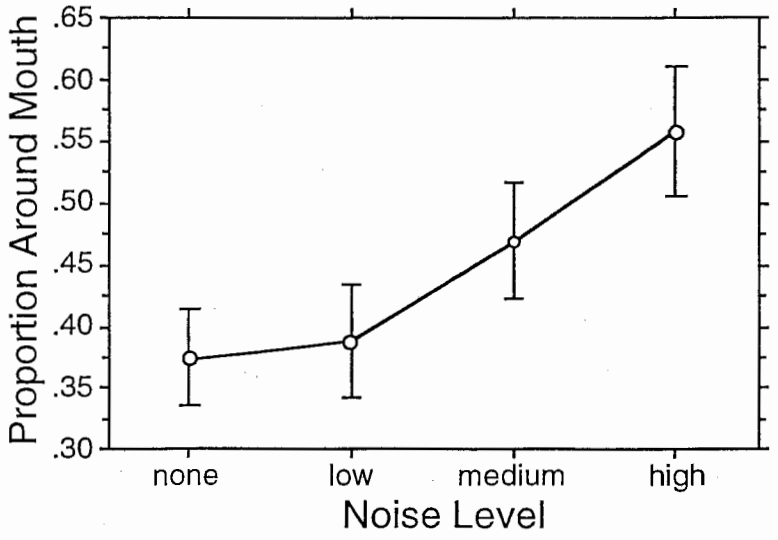

Figure 4. The proportion of fixations during a trial falling in the mouth bin is plotted as a function of masking noise level.

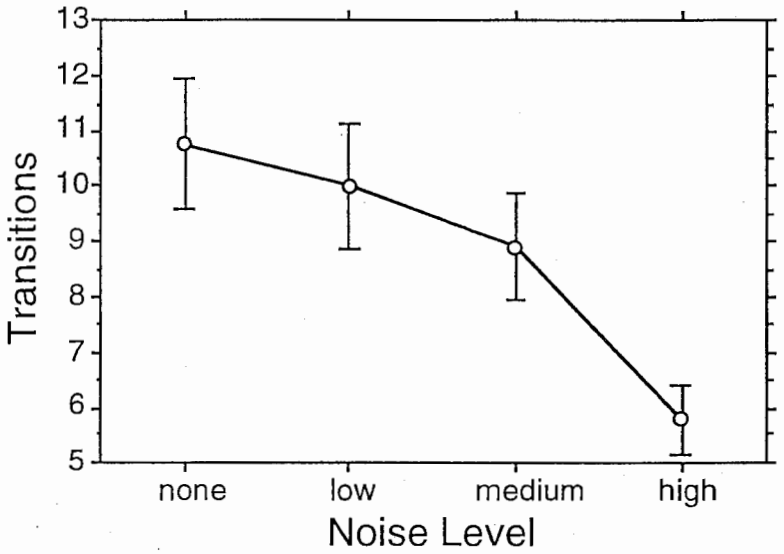

Figure 5. Means for the number of gaze transitions between eyes and mouth within a trial are plotted as a function of noise level.

When Do Perceivers Gaze At The Mouth?

In the following sub-sections, two analyses are presented which 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 perceivers shifted their gaze among the five facial regions is examined. Since this phenomenon has not previously been described for an audiovisual perception task, the aim here is to address several basic 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?

The eye position data were assigned to the five facial bins shown in Figure 3. However, in order to distinguish transitions from noise in the vicinity of the boundaries between bins, a minimum number of 5 consecutive samples of the same bin value was required for assignment to that bin. Using this constraint, the frequencies of all possible sequence patterns were computed for sequences containing 3, 4, 5, 6, and 7 gaze locations. The number of observable patterns for a given sequence length depends primarily on trial length and secondarily on the amount of eye motion noise at bin boundaries. As shown by the following relation, the number of possible pattern types (PT) depends on the sequence length (SL) and the number of analysis bins (B).

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

The resulting sets for sequence lengths 3 to 7 contain 80, 320, 1280, 5120 and 20,480 possible patterns, respectively. Since our purpose here was primarily descriptive and there were obvious differences between the two language groups, the data were separated accordingly. Table 1 gives the basic results for each language group as a function of sequence length. Pattern types and frequency counts are summarized for the overall data sets on the left side of the table, and totals for the subset of pattern types that accounted for 1 percent or more of the total are given on the right. The data for this subset of the corpus were further analyzed using Chi-Square tests for the effects of noise and image size.

As shown in the left half of Table 1, perceivers produced only small subsets of the possible patterns, and comparison of the language-specific with the "Combined" columns shows that the overlap in pattern sets used by the Japanese and English subjects was large. For example, for sequence length 3, a total of 58 pattern types were observed. As was typical throughout the corpus, English subjects produced a greater variety of pattern types than did Japanese subjects — 53 *vs.* 43. Only 5 pattern types were unique to the Japanese subjects (on the other hand, 15 pattern types were unique to the English subjects). The overlap was fairly stable across the different sequence lengths, while the language-specific variety of patterns increased to almost 2:1 for English *vs.* Japanese at sequence length 6.

The difference in the overall numbers of patterns (count) observed for English and Japanese subjects is consistent with the longer duration of stimuli for the English conditions. While the number of observed pattern types for both language groups increased with sequence length (and the total number possible), the subsets became proportionally smaller since the increase was basically linear rather than exponential. In

Table 1

Gaze Sequence Patterns and Pattern Counts As a Function of Gaze Sequence Length (SL) and Language (E, J).

		Overall				> 1%			
SL	Condition	E.	J	E+J	Possible	E	J	E+J	Total
3	patterns	53	43	58	80	12	12	12	
	count	3938	3227	7165		3560	3078	6638	93
4	patterns	132	88	154	320	23	21	24	
	count	3789	3070	6859		3331	2823	6154	90
5	patterns	252	143	300	1280	29	26	33	
	count	3647	2918	6565		2718	2355	5073	77
6	patterns	401	218	486	5120	27	20	27	
	count	3512	2770	6282		1883	1699	3582	57
7	patterns	564	327	680	20480	18	14	20	
	count	3384	2626	6010		1123	1251	2374	40

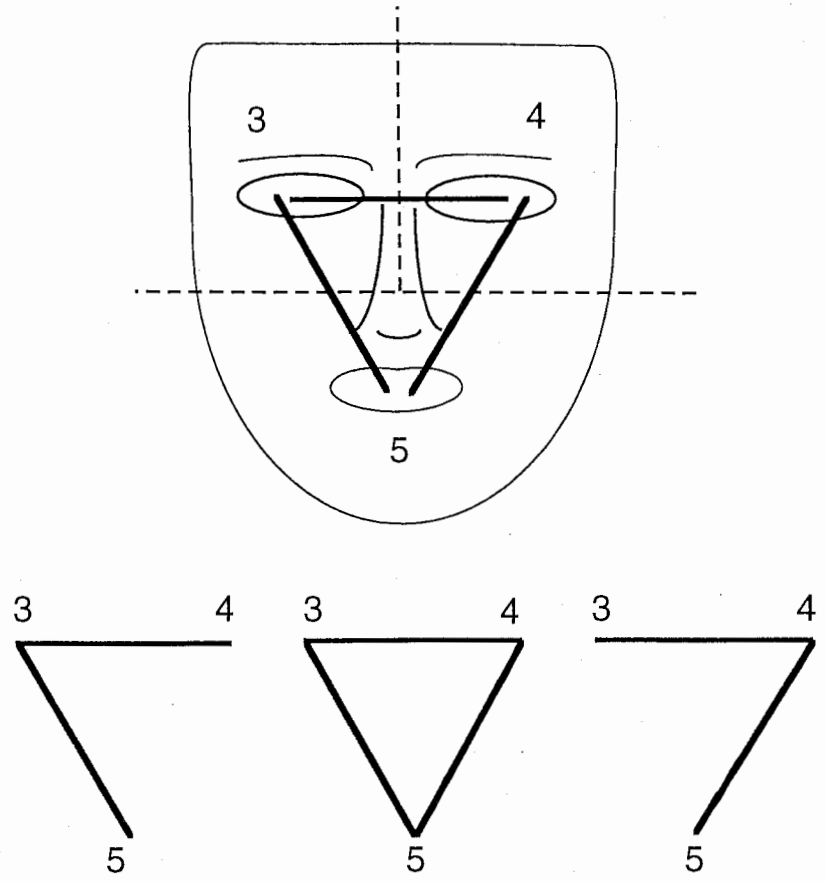

Figure 6. Overlaid on the schematic face are the most common gaze sequence patterns involving repetitive transitions between just two targets: eye-eye (3-4-3...) or eye-mouth-eye (3-5-3..., 4-5-4...). The next most common patterns are shown below: at the sides, the pattern templates for repetitive transitions between two targets with an occasional transitition 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 counter-clockwise sweeping of the three targets, e.g., 3-4-5-3-4.

large part, this can be attributed to the low incidence of patterns involving bins 1 and 2 above the eyes — less than 1 percent of the position data. That is, pattern sequences consisted primarily of the two eyes (3, 4) and the mouth (5), resulting in sequences such as 3-4-3, 4-5-4-3, 3-5-3-5-3, etc.

Indeed, were it not for the occasional instances where subjects produced patterns involving one of the forehead bins, we could recompute the equation above for just the three eye and mouth bins. The resulting number of possible patterns would then be 12, 24, 48, 96, 192 for sequence lengths 3 to 7, respectively. Consideration of the patterns accounting for 1 percent or more of the total observed, given on the right side of Table 1, would appear to justify such a recomputation since these "high frequency" patterns consisted entirely of gaze shifts between the eyes and mouth. The two shortest sequences, in particular, exhaustively exploit the recomputed set of possible patterns when results for the two language groups are combined. However, the number of observed pattern types reached a maximum of only 33 at sequence length 5, and there is less overlap in the pattern sets used by the two languages. Furthermore, the two longest sequence lengths showed progressively less pattern type variability.

The most common pattern types for sequence length 3 are schematized and ranked in Figure 6. This ranking persisted for all pattern lengths. That is, the most common patterns were repetitive sequences involving transitions between the two eyes (e.g., 3-4-3-4, 4-3-4-3-4-3), followed next by a pattern involving one of the eyes and the mouth (e.g., 3-5-3-

5, 5-3-5-3-5-3). 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. Next most common were patterns consisting of patterns combining a repetitive eye-eye sequence with a transition to the mouth and perhaps back again, e.g., 3-4-3-5, 4-5-4-3-4-3. The least common of the high frequency patterns were those tracing the apices of the eye-mouth triangle in a circular fashion, e.g., 3-4-5-3-4. English subjects consistently displayed a wider variety of patterns than did the Japanese for all sequence lengths except 3. Finally, the set of common patterns accounted for progressively less of the total corpus as sequence length increased — i.e., from 93 to 40 percent.

Table 2

Effects of Noise Level and Image Size Condition on Gaze Sequence Pattern Counts as a Function of Sequence Length (SL) and Language (E, J).

		Overall		> 1%		Overall			> 1%	
SL	Noise	E.	J	E	J	Image	E	J	E	J
3	None	1168	915	1119	845	1	967	630	891	603
	Low	1176	912	1039	894	2	812	977	812	971
	Mid	1035	773	919	755	3	1198	693	1109	693
	High	559	627	483	584	4	961	927	748	811
4	None	1128	875	1066	777	1	930	592	836	555
	Low	1138	873	968	831	2	774	937	766	903
	Mid	999	733	859	689	3	1160	654	1053	623
	High	524	589	438	526	4	925	887	676	742
5	None	1089	836	900	657	1	895	556	708	475
	Low	1103	835	804	689	2	739	898	644	751
	Mid	965	695	655	571	3	1122	616	878	501
	High	490	552	359	438	4	891	848	488	628
6	None	1051	799	655	485	1	862	520	504	364
	Low	1068	797	582	466	2	707	859	478	527
	Mid	935	658	391	435	3	1084	580	605	310
	High	458	516	255	313	4	859	811	296	498
7	None	1015	762	371	349	1	832	486	344	270
	Low	1034	760	332	337	2	679	821	276	377
	Mid	906	623	252	337	3	1046	545	366	196
	High	429	481	168	228	4	827	774	137	408

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. With regard to noise level effects, there were fewer patterns at higher noise levels for subjects of both languages. This agrees with the inference that gaze fixations are fewer and of longer duration at higher noise levels. Comparing the pattern counts representing 1% or more of the data for the two language groups, the differences between lower and higher noise levels is less extreme for the Japanese than English perceivers: At lower noise levels, fewer patterns are elicited for the Japanese; at higher noise levels, fewer patterns are 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 images sizes. Furthermore, the inconsistency differed for the two language groups. Thus, for the English group, the highest frequency of patterns was elicited for *size 3*, the next highest for *size 1*, the next for *size 2*, and the lowest frequency was observed for *size 4*. For the Japanese subjects, the high-to-low frequency counts were obtained for *size 2, 4, 3, 1* for sequence lengths 3-5, but *2, 4, 1, 3* for sequence lengths 6 and 7.

Table 3

Ranked Pattern Counts for Gaze Sequence Length Five by Condition — English.

	Noise Level				Image Size				
Sequence	None	Low	Mid	High	1	2	3	4	Total
43434	90	111	68	55	121	64	96	43	324
34343	88	110	53	60	116	64	84	47	311
45454	30	35	42	28	31	74	25	5	135
54545	28	31	33	26	22	72	16	8	118
43454	36	35	37	7	39	21	39	16	115
45434	35	33	37	5	33	25	42	10	110
35353	57	10	10	26	24	5	45	29	103
34345	32	31	25	12	21	25	36	18	100
53434	35	31	14	15	19	27	27	22	95
34543	36	28	28	2	24	21	34	15	94
54343	35	31	23	3	18	5	46	23	92
53535	45	13	10	24	23	19	35	15	92
54345	23	26	27	5	9	23	34	15	81
34534	21	25	26	9	26	20	28	7	81
34353	41	17	12	8	16	11	24	27	78
43435	31	27	10	6	17	14	19	24	74
35343	36	19	11	8	16	19	22	17	74
43534	28	22	11	5	12	14	19	21	66
45343	19	23	14	7	12	17	21	13	63
43543	19	24	16	3	12	14	22	14	62
43453	19	19	14	8	6	22	23	9	60
34545	10	13	21	11	18	11	19	7	55
53435	24	13	13	4	13	6	17	18	54
35434	21	20	10		11	12	19	9	51
34354	14	22	11	3	14	12	14	10	50
45345	10	7	26	6	5	9	25	10	49
54543	8	16	20	4	17	13	15	3	48
53534	20	4	11	7	9	3	13	17	42
53453	9	8	22	2	4	2	19	16	41
Total	900	804	655	359	708	644	878	488	2718

The pattern-specific frequencies for sequence length 5are tabulated for English and Japanese in Tables 3 and 4, respectively; the other four sequence lengths gave similar results and are tabulated in Appendix B. The trends in frequency were not uniform across the elicited pattern types. At higher noise levels, the highest frequency patterns, between the two eyes (e.g., 4-3-4-3-4), gave way somewhat to repetitive patterns between one eye and the mouth. Also the two language groups differed. For English perceivers, patterns involving the speaker's left eye (e.g., 5-4-5-4-5) became more frequent, particularly at the *middle* noise level. For Japanese subjects, the frequency of patterns involving the right eye and the mouth (e.g., 3-5-3-5-3) was about the same at the two higher noise levels as the repetitive eye-eye patterns.

In general, for both language groups, the diversity of high frequency patterns reduced as noise level increased. However, there was one striking difference between the two language groups in pattern diversity, common to all of the sequence lengths examined. For English subjects there was a large drop in frequency between the most frequent repetitive eye-eye patterns and the next most common pattern type, which initiated a gradual decline in frequency through the remaining pattern types. For Japanese subjects however, there was a relatively small frequency drop between the repetitive eye-eye pattern and the next most common repetitive right eye-mouth pattern, e.g., 3-5-3-5-3. There was then a large reduction in frequency between this and the third most common pattern, from which point pattern frequencies gradually declined.

As with the higher noise levels, 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), as well as a general increase in patterns involving at least one eye-mouth-eye circuit, e.g., 3-4-3-5-3.

Table 4

Ranked Pattern Counts for Gaze Sequence Length Five by Condition — Japanese.

	Noise Level				Image Size				
Sequence	None	Low	Mid	High	1	2	3	4	Total
43434	100	110	85	76	82	136	47	106	371
34343	94	106	89	78	76	138	49	104	367
35353	74	66	89	49	35	62	70	111	278
53535	73	61	81	46	24	56	68	113	261
45454	17	26	26	11	44	14	11	11	80
34345	27	20	17	15	16	31	17	15	79
43454	22	22	13	10	22	28	6	11	67
34353	21	22	12	9	10	24	17	13	64
54545	13	18	22	9	35	7	9	11	62
45434	21	22	10	6	23	18	9	9	59
53434	17	12	14	14	12	23	15	7	57
54343	21	15	8	7	14	16	13	8	51
43435	19	12	9	11	9	19	15	8	51
34543	19	16	8	4	16	14	7	10	47
35343	12	16	11	8	10	19	11	7	47
34534	13	13	8	12	6	14	13	13	46
53453	9	18	4	10	1	12	15	13	41
43534	13	16	4	7	6	18	12	4	40
53534	7	15	10	7	4	13	15	7	39
43535	11	10	14	4	3	13	10	13	39
43453	11	10	10	7	8	15	11	4	38
45353	10	13	6	8	2	10	16	9	37
45343	11	8	6	11	6	14	10	6	36
34545	6	14	7	7	7	14	8	5	34
35345	8	15	3	6	1	12	14	5	32
34535	8	13	5	6	3	11	13	5	32
Total	657	689	571	438	475	751	501	628	2355

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 ms (Calliard et al., 1996). Perceivers could conceivably make use of such phoneme-specific information to enhance perception.

The eye movement data for two subjects (see Footnote 2) were coded for target location and for the identity of the preceding, current, and following phonetic segment. A correlation with a temporal prior segment would suggest probabilistic prediction of visible oral events based presumably on a combination of prior acoustic and visual events, while correlations with the following segment would suggest a more simply reactive visual response. A correlation with the current segment could either imply prediction or pretuning of the oculomotor system to reduce reaction time (for review, see Carpenter, 1988) or a combination of predictive and reactive phenomena. However, no correlation between eye position and phonetic identity has been found for any of the comparsions 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 the 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, perceivers spent 45-70 percent of each trial gazing at the speaker's eyes, depending on the masking noise level. Also, 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.

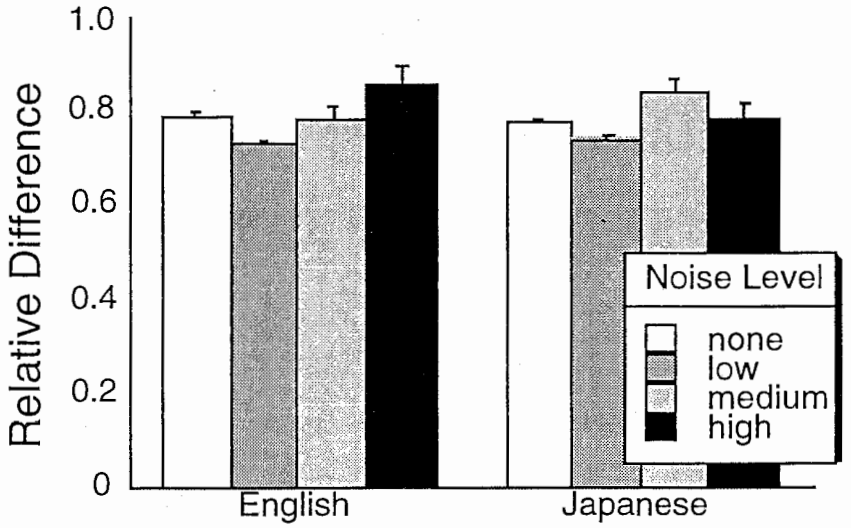

Figure 7. Shown is the proportion of fixations on the eyes within a trial attributable to one eye in particular. The relative difference is plotted as a function of noise level and language.

Eye position samples were assigned to bins *3* and *4* for the stimulus subject's right and left eye, respectively (see Figure 3). The values within each bin were summed and normalized for trial length differences. The proportion of each trial that 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 7.

The figure shows that perceivers gazed predominantly (> 70%) at only one of the eyes and that this asymmetrical preference increased by a few percent 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 where 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 that 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 is 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 perceivers used to shift their gaze to and from the mouth. In this section, we assess the effects of noise level and image size more locally by examining the eye movement behavior just in the vicinity of the mouth. The first measure examined assesses the variability of eye motion. In preliminary assessments of these data (e.g., Vatikiotis-Bateson et al., 1994), we 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 not so. 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 3). The per trial mouth 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, $\Sigma\mathbf{r}$, of all **r** for a given image-noise condition (i.e., across two trials) gives the average deviation from the centroid, $\underline{\mathbf{r}}$, when normalized for differences in trial length.

There was no reliable difference in centroid position for the condition-specific means. However, ANOVA on the range of motion for the mouth-centered data of nine 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 8.

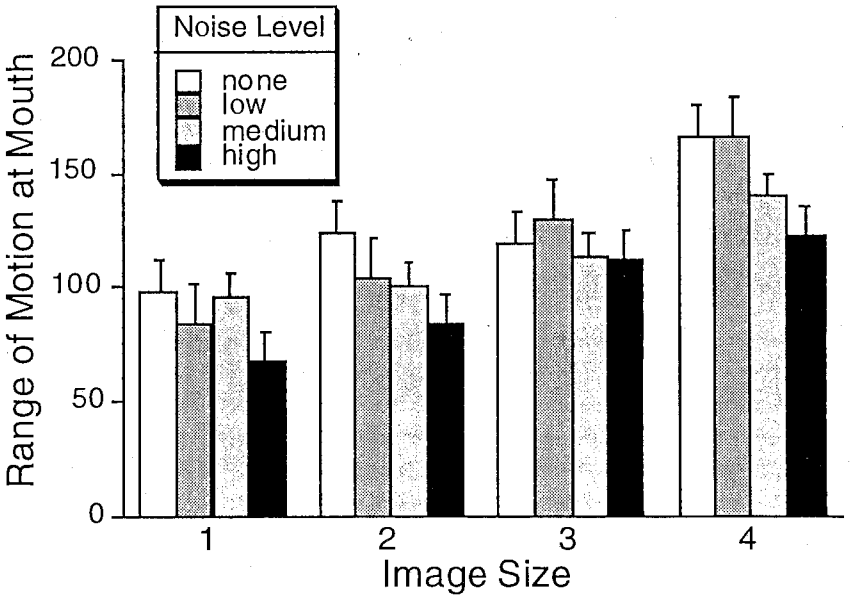

Figure 8. The main effects of noise level and image size on the range of motion around the mouth ($\underline{\mathbf{r}}$) are plotted together.

We believe the increase of $\underline{\mathbf{r}}$ at larger image sizes to be 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 to the overall variability. The reduced number of saccades between the eyes and mouth and longer fixation duration at higher noise levels, implies 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 Euclidian distance **r**, the ratio (**x/r**) was derived denoting the proportional contribution of horizontal motion. Analysis of variance 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 9, the proportion of the horizontal component, which is the larger of the two, decreased by about 6 percent as noise level increased.

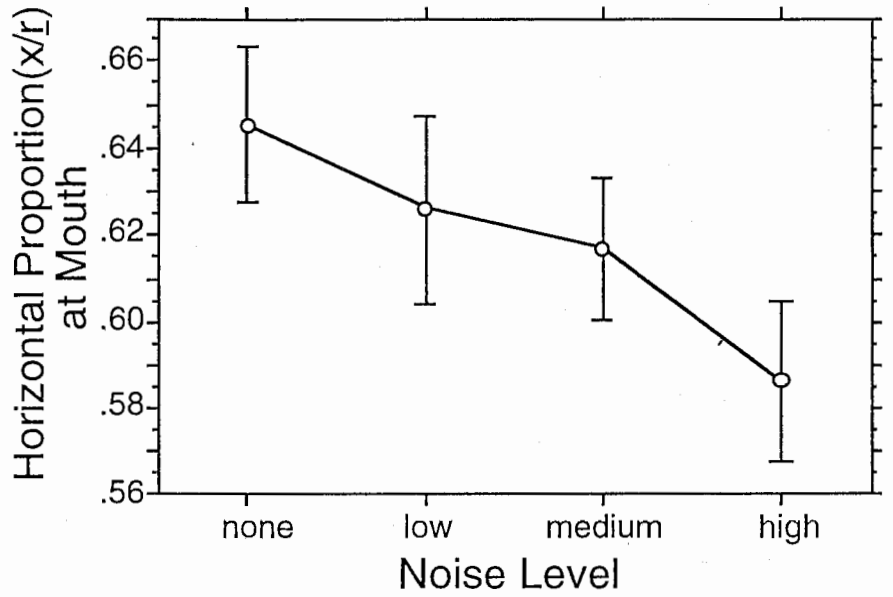

Figure 9. The proportional contribution, **x/r**, of the horizontal component (**x**) to the average distance from the centroid (**r**)is plotted as a function of masking noise level for the mouth region.

This result is interesting because the overall proportionality of the horizontal and vertical components reflects the geometry of the mouth target — i.e., the mouth is about twice as wide as it is high, or 67 and 33 percent, 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 than the horizontal component. This problem is addressed by a third analysis which provides rudimentary evidence that transitions to the mouth from one eye arrive in different areas of the horizontally arrayed mouth target than transitions from the other eye.

The *mouth* bin (see Figure 3) was divided into two halves along the vertical midline of the lips.[3] Then, absolute relative differences in the proportion of position data falling on one side of the mouth or the other were computed. Similar to perceivers' preference to fixate on one eye, a preference was found for one side of the speaker's mouth. 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, eye-mouth transitions usually crossed the vertical midline rather than remain on the same side, as shown by comparison of opposite (.55) and same side (.25) correlations between *eye* bins and the subdivided *mouth* bin. That is, a transition from the left 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.

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 methodolgical 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, subjects said they preferred the two middle image sizes over the smallest and largest image sizes. Based on this preference, we have used image size 3 in subsequent studies (e.g., Appendix A; 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 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 more than 10.5 degrees, so 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.

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 perceivers to fixate on the speaker's lips. At the lowest noise levels, fixations on the eyes accounted for more than 65 percent 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 percent of the trial. We had assumed the primary function of interlocutor eye contact to be sociolinguistic, mediating turn-taking, sincerity, etc. In a non-interactive experimental context such as this, the sociolinguistic factors attendant to eye contact should be largely irrelevant. We thought it much more likely that subjects would retrieve perceptually relevant visual correlates to the prosody and phonetics from the facial deformation patterns associated with jaw and lip motion. So, why did subjects persist in fixating so much on the eyes at the highest noise levels? Why not simply fixate on the region around the mouth? There are a number of possible answers we would like to see investigated further.

One possibility, which is discussed further in the next section, is that eye contact serves some fundamental, structural role in the retrieval of visual information from the vicinity of the mouth. The intelligibility results support this to the extent that, at higher noise levels where visual information becomes most critical, the integrity of the visual information began to break down when the angle between eyes and mouth reached 8-9 degrees. The only indication of an effort by subjects to compensate for this difficulty was in the noise level and image size effects on the frequency of specific patterns of saccadic gaze shifts between the eye and mouth targets. Since the subjects had no experience with the task prior to the study, we do not know whether or not subject performance under similar noise level and image size conditions would improve if given more time. For example, perceivers might further adapt their eye movement behavior to optimize intelligibility, or they might become more adept at retrieving relevant visual information using the same behavioral strategies — i.e., strategies entailing substantial fixations on the eyes, but with better retrieval from the visual near periphery.

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 cannot report exactly how consistent subjects were in their choice of preferred eye across trials. However, it is clear from even a cursory examination of the data that perceivers do vary their choice of preferred eye across trials. Since there was also a strong preference for a particular eye in the repetitive eye-mouth gaze sequence patterns, suggesting the possibility that the preferred eye acts 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 ms (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 in particular are worthy of mention because we have been able to address them in a subsequent study. First, the strictly phonetic bias of the post-trial 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 to the phonetic detail of every word may be quite abnormal. Second, the finding reported in Appendix A 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. Both of these factors have been tested in a subsequent study. Preliminary results (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, phonetic 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 Audiovisual Speech

The major implication of the results is that perceiver eye movement behavior — a largely voluntary, motor event — does have a role in audiovisual speech perception. Of course, the exact contribution of perceiver eye motion to speech perception remains to be discovered. Furthermore, the fact that perceivers adapted their eye movement behavior to changes primarily in the acoustic, rather than the visual, environment points up the need to determine what other concurrent 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). A useful precursor to the tasks of distinguishing the morphological and task-specific influences on eye behavior and the contribution of eye movement control on speech perception 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 production of audiovisual behavior (e.g., Vatikiotis-Bateson et al., 1996; for review, see Munhall and Vatikiotis-Bateson, in press).

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 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 is where the phonetic information is primarily located. Yet, assuming some phonemes are more visually salient than others, the lack of correlation between gaze location and serial phonetic structure suggests that eye motion *per se* was not used to facilitate the perception of specific phonemes.

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. Although certain extreme phonetic postures might be identified from static images — e.g., pursed lips for a bilabial plosive (/p, b, m/), lip rounding (/u, o/) or spreading (/i, s/), visual information must be dynamic in order to enhance phonetic perception substantially. For example, Viktovich 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 and colleagues (Remez. Rubin, Berns, Pardo, & Lang, 1994) have demonstrated the dynamic nature of speech perception by using sine wave re-synthesis to remove the acoustic complexity of the speech signal while retaining its fine-grained temporal structure.

Thus, by foveating primarily on the eyes in audiovisual perception, perioral events might be more accurately detected by the temporally acute near periphery. 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.[4] 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 and, therefore, predictable events.

This account of the role of the near periphery on the extraction of phonetic information 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 therefore shapes the acoustic output), simultaneously produce dynamic deformations of the entire face.

Recently, the physiology and kinematics of facial motion during production of realistic speech has been examined through analysis of muscle EMG, facial kinematics, and the speech acoustics (Vatikiotis-Bateson et al., 1996; Vatikiotis-Bateson & Yehia, 1996a). 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 percent 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 offers largely redundant motion information. Second, the RMS amplitude of the acoustics is equally well recovered (80%) from either perioral or more remote facial regions, but is better recovered (89%) by combining the two regions, suggesting that correlates of the segmental acoustics are distributed non 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., Benoit et al., 1992). 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 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.[5]

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 insure slight differences in the actual facial deformation, we suggest that these are exactly the sort of differences to which perceivers rapidly adapt in audiovisual 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 produce habituated eye movement patterns that serve both phonetic and higher level, sociolinguistic criteria. While these patterns may be sufficient in a wide range of environments, including highly artificial perception tasks such as ours, they may be difficult to change.

The Effects Of Language On Eye Movement Patterning

Coherent language-specific differences in eye movement behavior appeared only in the analysis of gaze sequences. First, for both groups, the most frequent pattern type after the repetitive eye-eye pattern, was a repetitive eye-mouth pattern. However, the Japanese predominantly chose the right eye, while the English chose the left eye. At this point, we have no explanation for this difference; it could be due to a difference between the two stimulus speakers, such as the amount of head motion or expressive gestures. Second, English perceivers used a greater variety of gaze sequence patterns than did the Japanese, and at higher noise levels the tendency to reduce the variety of patterns was more pronounced for the Japanese group. This may be due to a combination of cultural and linguistic differences in the style and utility of gaze strategies for the two language groups.

It is often remarked that Japanese interlocutors tend to avoid direct eye contact, while westerners get nervous if it is absent. Although formal investigations of this phenomenon appear to be lacking, it is easily observed that there are indeed many situations, apparently prescribed by social status, gender, and probably other factors, where mutual eye contact will not be made by Japanese interlocutors. However, the failure to achieve mutual eye contact does not mean that Japanese interlocutors do not watch each other's faces and retrieve sociolinguistic or even phonetic information. Typically, one interlocutor will gaze directly at the face of the other, while the other looks elsewhere. Sometimes, the listener watches the talker's face while the talker looks elsewhere; other times, the situation is reversed. That is, face-to-face communication among Japanese interlocutors provides ample opportunity for audiovisual perception. Only the emphasis on making mutual eye contact is lacking. Perhaps, the greater variability in gaze sequence patterns for English perceivers reflects this more demanding sociolinguistic constraint on mutual eye contact among interlocutors, something independent of a strategy for enhancing audiovisual perception. That is, the greater pattern variability did not lead to better intelligibilty results for English subjects.

There is however the further possibility that Japanese and English speakers impart different degrees of linguistically relevant visual information. This was pointed out by Sekiyama and Tohkura (1993), who compared McGurk-style audiovisual tests for English and Japanese. They reported 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). They suggested that the Japanese phonetic system provides fewer salient visual correlates than English, which would lead to fewer mismatches in a McGurk paradigm. An important implication of the Sekiyama and Tohkura study for the results of the current 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 the McGurk study.

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 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 current study has set out a methodological and conceptual framework for pursuing a better understanding of audiovisual perception and its relation with speech production.

Acknowledgment

We thank 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 de Universität München for generous support and discussion during a phase of the writing.

References

Abry, C., Lallouache, T., & Cathiard, M. (1996). How can coarticulation models account for speech sensitivity to audio-visual desynchronization? In D. Stork and M. Hennecke (Eds.), *Speechreading by Humans and Machines.* (NATO-ASI Series F: Computer and Systems Sciences, vol. 150, pp. 461-471). Berlin: Springer-Verlag.

Benoit, C., Lallouache, T., Mohamadi, T., & Abry, C. (1992). A set of French *visemes* for visual speech synthesis. In G. Bailly, C. Benoit, & T.R. Sawalis (Eds.), *Talking Machines: Theories, Models, and Designs.* pp. 335-348. Amsterdam: Elsevier Science Publishers.

Bertelson, P., & Radeau, M. (1976). Ventriloquism, sensory interaction and response bias: Remarks on the paper by Choe, Welch, Gilford, and Juola. *Perception & Psychophysics*, **19**(6), 531-535.

Carpenter, R.H.S. (1988). *Movements of the Eyes.* London: Pion Limited (2nd revised ed.).

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

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.

Leigh, R.J. & Zee, D.S. (1991). Oculomotor disorders. In R.H.S Carpenter (Ed.), *Eye movements.* Vol. 8 in *Vision and visual disfunction* series. London: Macmillan Press.

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). New Jersey: Lawrence Erlbaum & Associates.

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

Gruesser, O.J. & Landis, T. (1991). *Visual agnosias and other disturbances of visual perception and cognition.* Vol. 12 in *Vision and visual disfunction* series. London: Macmillan Press.

Hirayama, M., Vatikiotis-Bateson, E., Gracco, V., & Kawato, M. (1994). Neural network prediction of lip shape from muscle EMG in Japanese speech. In *The 1994*

International Conference on Spoken Language Processing (ICSLP-94), **2** (pp. 587-590). Yokohama, Japan.

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: C.J. Hogrefe.

Johnson, J. A., Rosenblum, L. D., & Saldaña, H. M. (1994). The contribution of a reduced visual image to speech perception in noise. *Journal of the Acoustical Society of America*, **95**(No. 5, Pt. 2), 3009.

Kurita, T., Honda, K., Kakita, Y. (1994). A physiological model for the synthesis of lip articulation. *Journal of the Acoustical Society of Japan,* **50**, 465-473 (in Japanese).

Lansing, C. R., & McConkie, G. (1995). A new method for speechreading research: Tracking observer's eye movements. *Journal of the Academy of Rehabilative Audiology*, **27**, 25-43.

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, **21**, 131-141.

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.

Mase, K. & Pentland, A. (1991). Automatic lipreading by optical-flow analysis. *Systems and Computers in Japan,* **22**, 67-75.

Massaro, D. (1987). *Speech Perception by Ear and Eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Moray, N. (1993). Designing for attention. In *Attention: Selection, awareness, and control, a tribute to Donald Broadbent* (pp. 53-72). Oxford: Clarendon Press.

Munhall, K. G., & Vatikiotis-Bateson, E. (in press). The moving face during speech communication. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by Eye, Part 2: The Psychology of Speechreading and audiovisual speech.* London: Taylor & Francis - Psychology Press.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, **58**(3), 351-362.

Poizner, H., Bellugi, U., & Klima, E. (1990). Biological foundations of language: Clues from sign language. *Annual Review of Neuroscience*, **13**, 283-307.

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

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

Reisberg, D., McLean, J., and Goldfield, A. (1987). Easy to hear but hard to understand. In B. Dodd and R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lipreading*, pp. 97-114. New Jersey: Lawrence Erlbaum Associates.

Schiffman, H. R. (1982). *Sensation and perception: An integrated approach* (2nd ed.). New York: John Wiley & Sons.

Sekiyama, K., and Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *J. of Phonetics*, **21**, 427-444.

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, Series F: Computers and Systems Sciences) *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 and R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lipreading*, pp. 3-52. New Jersey: Lawrence Erlbaum Assoc.

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

Vatikiotis-Bateson, E., Munhall, K. G., Hirayama, M., Kasahara, Y., & Yehia, H. (1996). Phsyiology-based synthesis of audiovisual speech. In *4th Speech Production Seminar: Models and Data,* (pp. 241-244). Autrans, France:.

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) *150* (pp. 221-232). Berlin: Springer-Verlag.

Vatikiotis-Bateson, E., & Yehia, H. C. (1996). Physiological modeling of facial motion during speech. *Trans. Tech. Com. Psycho. Physio. Acoust.*, **H-96**(65), 1-8.

Vatikiotis-Bateson, E., & Yehia, H. C. (1996b). Synthesizing audiovisual speech from physiological signals. In *The 3rd Joint Meeting of the Acoustical Societies of America and Japan,* 2-6 December, 1996, Honolulu, HI: in press.

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

Waters, K. (1987). A muscle model for animating three-dimensional facial expression. *Computer Graphics,* **21**, 17-24.

Waters, K. & Terzopoulos, D. (1992). The computer synthesis of expressive faces. *Philosophical Transactions of the Royal Society of London B*, **335**, 87-93.

Wolff, G. J., Prasad, K. V., Stork, D. G., & Hennecke, M. (1994). Lipreading by nerual 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. *ITE Tech Rep.,* **17,** 1-10. (in Japanese).

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

Zangemeister, W., & Stark, L. (1981). Active head rotations and eye-head coordination. *Annals of the New York Academy of Science*, **374**, 540-549.

Appendix A

A perception study was conducted to test the possibility that the stimulus monologues differed in their inherent intelligibility due, for example, to differences in syntactic structure, length, or lexical-semantic content. Forty 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 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 five. 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 three.

The effects of monologue and presentation order on intelligibility were analyzed using 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, intelligibility scores for two or three of the monologue pairs were markedly different from the mean. 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 (see Table A1). Also, a third monologue in the Japanese series had a much lower than average intelligibility score. As this case corresponded to a *no noise* 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.

Table A-1

Mean Monologue Intelligibility Scores (0-1) by Language, Noise Level and Image Size.

		English				Japanese			
		Size 1	Size 2	Size 3	Size 4	Size 1	Size 2	Size 3	Size 4
None	M	0.43	0.5	0.52	0.51	0.42	0.62	0.66	0.62
	SEM	0.05	0.04	0.06	0.06	0.06	0.04	0.04	0.04
Low	M	0.49	0.52	0.41	0.48	0.51	0.68	0.67	0.69
	SEM	0.04	0.04	0.06	0.05	0.05	0.04	0.04	0.04
Mid	M	0.48	0.43	0.45	0.44	0.62	0.67	0.83	0.44
	SEM	0.05	0.04	0.05	0.04	0.04	0.04	0.03	0.04
High	M	0.25	0.57	0.64	0.43	0.66	0.72	0.65	0.63
	SEM	0.04	0.04	0.06	0.05	0.05	0.04	0.04	0.03

Note — The results are tabulated here to reflect the way the monologues were assigned to noise level and image size conditions in the production study.

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 interacttion 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 percent. 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 100-130 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 2), and the small perceptual variability among monologues when tested at a consistent noise level, suggest that masking noise level was the primary determinant of intelligibility in the production study.

Appendix B

Tables are presented below for sequence lengths 3, 4, 6, and 7. Gaze sequence patterns are ranked by frequency for each language group as a function of noise level on the left and image size on the right.

Table B-1

Ranked Pattern Counts for Gaze Sequence Length Three by Language and Condition.

	Noise Level				Image Size				
Sequence	None	Low	Mid	High	1	2	3	4	Total
				English					
434	187	215	142	87	205	136	197	93	631
343	188	200	117	95	189	119	177	115	600
454	90	99	116	48	88	134	91	40	353
345	78	75	92	28	59	61	101	52	273
353	121	47	43	53	55	28	89	92	264
534	83	66	76	31	46	47	86	77	256
543	78	87	77	13	65	55	84	51	255
545	56	67	83	45	56	113	52	30	251
435	77	61	44	18	45	40	58	57	200
535	83	28	32	41	35	17	73	59	184
453	41	44	59	17	20	34	62	45	161
354	37	50	38	7	28	28	39	37	132
Total	1119	1039	919	483	891	812	1109	748	3560
				Japanese					
434	168	166	127	118	135	209	87	148	579
343	161	164	131	123	120	209	97	153	579
353	118	117	130	76	58	111	122	150	441
535	102	98	118	71	39	95	108	147	389
454	49	61	51	29	76	54	31	29	190
345	52	60	31	31	35	59	44	36	174
534	39	55	28	33	20	51	49	35	155
545	28	43	41	26	52	34	30	22	138
435	39	37	28	20	17	43	35	29	124
453	30	40	22	30	12	41	41	28	122
543	39	38	29	15	31	43	28	19	121
354	20	15	19	12	8	22	21	15	66
Total	845	894	755	584	603	971	693	811	3078

Table B-2

Ranked Pattern Counts for Gaze Sequence Length Four by Language and Condition.

	Noise Level				Image Size				
Sequence	None	Low	Mid	High	1	2	3	4	Total
				English					
3434	125	148	86	76	147	92	128	68	435
4343	125	147	79	66	144	85	124	64	417
4545	40	47	59	40	42	92	36	16	186
4345	55	57	52	17	47	45	64	25	181
5454	40	50	57	32	41	92	34	12	179
5434	59	58	52	8	50	41	63	23	177
5343	60	47	32	21	34	33	47	46	160
3454	48	42	51	13	45	32	53	24	154
4543	44	47	51	6	41	34	52	21	148
5353	66	17	22	33	27	8	59	44	138
3535	68	17	15	34	30	10	56	38	134
3435	58	40	23	12	31	25	40	37	133
3534	49	29	25	12	23	18	32	42	115
4534	29	31	41	13	17	27	47	23	114
4353	50	26	17	13	23	17	28	38	106
3453	29	29	37	11	11	26	43	26	106
3543	29	35	20	4	18	17	29	24	88
4354	23	34	26	3	21	20	27	18	86
5345	21	15	39	9	10	13	35	26	84
5435	18	21	21	4	14	14	17	19	64
5453	10	11	20	5	8	8	16	14	46
4535	12	8	16	4	2	6	14	18	40
3545	8	12	18	2	10	11	9	10	40
Total	1066	968	859	438	836	766	1053	676	3331
				Japanese					
3434	128	132	108	102	101	173	69	127	470
4343	121	125	99	90	93	159	64	119	435
5353	88	84	102	59	37	74	94	128	333
3535	86	73	101	54	35	72	80	127	314
4345	35	33	24	17	31	43	19	16	109
4545	21	32	32	17	45	23	18	16	102
5454	19	30	31	15	48	20	15	12	95
5343	25	27	20	19	18	35	23	15	91
3435	29	28	18	15	17	30	23	20	90
3454	28	31	16	13	24	32	15	17	88
4353	27	29	20	12	10	33	25	20	88
5434	31	28	17	9	29	30	16	10	85
3453	22	28	14	18	10	27	27	18	82
3534	20	32	14	15	11	31	27	12	81
4543	23	27	15	8	25	26	12	10	73
4534	16	18	11	17	7	18	19	18	62
5345	14	26	7	13	2	16	23	19	60
4535	12	20	9	12	4	20	19	10	53
5453	8	12	8	10	2	14	14	8	38
3543	14	6	12	5	4	14	10	9	37
5354	10	10	11	6	2	13	11	11	37
Total	777	831	689	526	555	903	623	742	2823

Table B-3

Ranked Pattern Counts for Gaze Sequence Length Six by Language and Condition.

	Noise Level				Image Size				
Sequence	None	Low	Mid	High	1	2	3	4	Total
				English					
343434	65	85	47	51	98	52	66	32	248
434343	65	84	44	46	97	48	64	30	239
545454	21	26	29	20	20	60	13	3	96
454545	21	21	26	23	17	59	12	3	91
434543	29	24	21	2	20	15	30	11	76
535353	39	9	5	20	13	4	39	17	73
345434	28	21	22	1	19	15	29	9	72
353535	38	8	4	18	17	2	36	13	68
543434	24	23	18	1	14	14	27	11	66
434345	23	20	18	5	18	11	27	10	66
534343	23	21	7	8	12	15	17	15	59
343454	17	18	19	4	17	10	19	12	58
543454	19	17	18	3	22	11	20	4	57
454345	16	17	19	2	14	11	22	7	54
454343	19	16	16	3	19	12	20	3	54
343534	24	15	7	4	8	10	18	14	50
434534	13	17	12	7	5	19	19	6	49
343435	20	19	5	4	12	10	14	12	48
345343	16	19	7	5	5	16	16	10	47
435343	22	15	6	3	10	10	14	12	46
353434	20	13	3	4	4	14	16	6	40
343453	15	12	6	7	11	12	9	8	40
435434	15	16	8		7	11	14	7	39
534353	21	6	7	3	4	14	11	8	37
453434	13	14	5	5	6	4	12	15	37
343543	12	15	7	3	7	12	10	8	37
434353	17	11	5	3	8	7	11	10	36
Total	655	582	391	255	504	478	605	296	1883
				Japanese					
343434	80	93	76	66	67	116	38	94	315
434343	74	87	70	61	61	108	34	89	292
535353	65	54	71	42	24	48	60	100	232
353535	61	50	73	40	24	48	52	100	224
434345	22	17	13	10	14	25	11	12	62
545454	12	15	20	6	35	4	6	8	53
454545	12	15	18	7	32	3	7	10	52
343454	17	11	8	10	9	21	5	11	46
345434	17	15	7	3	16	11	6	9	42
343435	13	10	9	10	8	18	9	7	42
543434	16	13	7	5	12	15	8	6	41
534343	11	11	10	8	8	17	10	5	40
434543	15	12	7	3	14	12	2	9	37
434353	14	10	7	6	5	15	11	6	37
454343	12	10	6	5	10	9	7	7	33
343534	12	12	3	6	6	14	9	4	33
435353	7	10	12	3	3	11	8	10	32
345343	10	6	5	8	5	11	9	4	29
343453	8	8	8	5	6	10	10	3	29
453434	7	7	5	9	5	11	8	4	28
Total	485	466	435	313	364	527	310	498	1699

Table B-4

Ranked Pattern Counts for Gaze Sequence Length Seven by Language and Condition.

	Noise Level				Image Size				
Sequence	None	Low	Mid	High	1	2	3	4	Total
				English					
4343434	52	68	39	39	84	40	52	22	198
3434343	53	69	30	43	81	43	47	24	195
4545454	16	18	22	18	16	49	9		74
5454545	16	19	21	15	11	51	7	2	71
4345434	24	19	18	1	17	12	26	7	62
3535353	32	6	2	14	12	2	30	10	54
5353535	27	7	3	15	10	2	32	8	52
4343454	13	13	14	1	12	6	18	5	41
4543434	12	14	14	1	12	5	16	8	41
5343434	13	13	6	8	7	11	12	10	40
4543454	14	10	13	2	13	8	15	3	39
5434543	16	11	12		17	6	15	1	39
3454343	14	10	13	1	10	8	14	6	38
3434543	13	13	9	2	7	8	13	9	37
3435343	20	10	4	3	7	7	15	8	37
3434345	11	9	12	4	7	7	16	6	36
5434343	11	12	11	1	12	4	14	5	35
3454345	14	11	9		9	7	15	3	34
Total	371	332	252	168	344	276	366	137	1123
				Japanese					
4343434	65	79	61	52	55	93	27	82	257
3434343	57	72	61	53	48	90	27	78	243
3535353	56	43	64	36	24	42	45	88	199
5353535	54	42	59	35	17	38	44	91	190
3434345	19	15	13	8	12	23	9	11	55
4545454	12	13	17	4	32	2	5	7	46
5454545	9	12	14	4	26	1	5	7	39
4343454	15	9	8	7	9	18	3	9	39
5434343	13	11	7	4	10	13	6	6	35
4345434	13	11	6	3	14	10	1	8	33
3434353	10	8	7	6	5	14	7	5	31
3434543	11	9	6	3	7	11	2	9	29
5343434	7	8	8	5	6	11	8	3	28
4343435	8	5	6	8	5	11	7	4	27
Total	349	337	337	228	270	377	196	408	1251

Footnotes

[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 meters. In order to increase the naturalness of the monologue presentation, speakers in a subsequent study (Eigsti et al., 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] Due to an oversight, masking noise and speech were mixed on both audio tracks for eight subjects. Only for the last two subjects, where the two tracks were separated and mixed only for loudspeaker presentation, could the clear speech channel be digitized simultaneously with eye movement.

[3] 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 and 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 $\pm.1$; fixations on the corners of the mouth were assigned $\pm.5$; 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 10, 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 towards the left corner of the mouth as noise level increased. English speakers moved towards the right, but the trend was not consistent for the highest noise level.

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

[5] 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 is further diffused once the damped connective fascia and relatively stiff outer skin layers are considered.