Published in final edited form as:
Brain Res. 2008 November 25; 1242: 162–171. doi:10.1016/j.brainres.2008.06.083.

The Effect of Varying Talker Identity and Listening Conditions on Gaze Behavior During Audiovisual Speech Perception

Julie N. Buchan[1], **Martin Paré**[1,2], and **Kevin G. Munhall**[1,3]

1*Department of Psychology, Queen's University, Kingston, Ontario, Canada*

2*Department of Physiology, Queen's University, Kingston, Ontario, Canada*

3*Department of Otolaryngology, Queen's University, Kingston, Ontario, Canada*

Abstract

During face-to-face conversation the face provides auditory and visual linguistic information, and also conveys information about the identity of the speaker. This study investigated behavioral strategies involved in gathering visual information while watching talking faces. The effects of varying talker identity and varying the intelligibility of speech (by adding acoustic noise) on gaze behavior were measured with an eyetracker. Varying the intelligibility of the speech by adding noise had a noticeable effect on the location and duration of fixations. When noise was present subjects adopted a vantage point that was more centralized on the face by reducing the frequency of the fixations on the eyes and mouth and lengthening the duration of their gaze fixations on the nose and mouth. Varying talker identity resulted in a more modest change in gaze behavior that was modulated by the intelligibility of the speech. Although subjects generally used similar strategies to extract visual information in both talker variability conditions, when noise was absent there were more fixations on the mouth when viewing a different talker every trial as opposed to the same talker every trial. These findings provide a useful baseline for studies examining gaze behavior during audiovisual speech perception and perception of dynamic faces.

1.0 Introduction

We see and process faces every day in a wide variety of contexts, from line drawings of faces and static photographs, to dynamic movies and live faces during face-to-face communication. These faces contain a wealth of social, emotional, identity and linguistic information. Although a great deal of information can be gleaned from static faces, the motion of dynamic faces contains information about identity and emotion not present in static faces (Ambadar, Schooler & Cohn, 2005; Hill & Johnson, 2001; Knappmeyer, Thornton & Bülthoff, 2003; O'Toole, Roark, & Abdi, 2002; Lander & Bruce, 2000). Facial motion also contains linguistic information, as evidenced by the fact that silent speechreading is possible (Bernstein, Demorest & Tucker, 2000). Rarely though, is this visual speech information present in the complete absence of auditory speech information and audiovisual speech perception is the natural manner of communication. Visual speech information influences the perception of auditory speech in both perfectly audible conditions (McGurk & MacDonald, 1976, MacDonald &

Correspondence: J.N. Buchan, Department of Psychology, Queen's University, Humphrey Hall, 62 Arch Street, Kingston, Ontario, Canada K7L 3N6, Tel: 613-533-6275, E-mail: 2jnb@queensu.ca.

McGurk, 1978) and acoustically degraded conditions (Sumby & Pollack, 1954; Erber, 1969; O'Neill, 1954). This audiovisual speech perception in face-to-face communication is one of the most naturally occurring, but perhaps least understood, instances of multisensory integration.

The visual information present in dynamic faces is constantly changing, and as such, we must adopt perceptual strategies that allow us to efficiently sample the changing visual information when it is available. One valuable approach to investigating how such visual information is gathered is by examining what facial features are being preferentially selected for more detailed processing by gaze fixations. Studies of gaze fixations have helped our understanding of face learning (Henderson, Williams & Falk, 2005), face recognition (Stacey, Walker & Underwood, 2005; Barton, Radcliffe, Cherkasova, Edelman & Intriligator, 2006; Althoff & Cohen, 1999; Hsiao & Cottrell, 2007), and social perception (Yarbus, 1967; Walker-Smith, Gale, & Findlay, 1977). Although the location of gaze fixations is influenced by low-level images properties of the stimuli (e.g. color, spatial frequency : Parkhurst, Law, Niebur, 2002; Parkhurst & Niebur, 2003), the locations chosen for selective visual processing are also knowledge driven (Henderson, 2003). The knowledge driven nature of gaze fixations is also evident in the gaze behaviors exhibited while watching talking faces. Gaze behavior studies examining audiovisual (Buchan, Paré & Munhall, 2007) and visual–only (Lansing & McConkie, 1999) speech have shown that task instructions influence facial locations from which information is gathered.

The influence of task instructions on gaze behavior, however, is surprisingly modest (Buchan, Paré & Munhall, 2007; Lansing & McConkie, 1999). While task instructions cause a slight shift in the preferred fixation locations on the face, the overall spatial distribution of fixations nonetheless remains fairly similar across tasks. One possible explanation for this similarity across tasks is that the gaze behavior exhibited during speech perception experiments is part of a face information gathering routine in which perceivers have other concurrent goals besides simply perceiving speech. Even in a context where subjects are watching videos for a speech experiment they may be sampling emotional and identity information, as well as social information such as the attention and intention of the speaker (Emery 2000; Baron-Cohen et al., 2001). This would provide an account for the considerable fixations made to the eyes during audiovisual speech experiments (e.g. Vatikiotis-Bateson, Eigsti, Yano & Munhall, 1998; Paré, Richler, ten Hove & Munhall, 2003; Lansing & McConkie, 2003) even though the eyes are not a particularly informative facial feature for perceiving speech sounds.

Identity information is present in both the face and the voice and recently there has been considerable interest in the ability of both humans and animals to use this information (e.g., Ghazanfar et al., 2007). Individuals can estimate body size, age and gender from both faces and voices. There are also strong correlations between the voice and the dynamic face and thus it is not surprising that identity can be matched cross-modally (Lachs & Pisoni, 2004; Kamachi, Hill, Lander & Vatikiotis-Bateson, 2003). Beyond the obvious use in recognizing individuals, voices and faces contain indexical information that influences other tasks such as speech perception (Nygaard, Sommers & Pisoni, 1994; Nygaard & Pisoni, 1998; Yakel, Rosenblum, & Fortier, 2000). For example, talker familiarity improves acoustic speech identification in noise (Nygaard et al., 1994) and it improves silent lip-reading (Yakel et al., 2000).

The influence of talker familiarity in speech perception raises the issue of whether the various dimensions of information visible on the face (identity, emotion, speech, etc.) always interact. In a study using a speeded classification task of static identity, emotion and speech images, it has been shown that when identity judgment is the primary task there is no interference from irrelevant variability in emotional expression or facial speech postures (e.g., Schweinberger & Soukup, 1998). However, when either emotion or speech perception was the primary task,

irrelevant variability in identity did interfere with reaction times in the primary task. It is still unknown, however, how identity and speech processing might jointly determine gaze behavior in dynamic stimuli.

In order to examine the influence of one facial dimension on the processing of another we manipulated the variability in the identity of talking faces during a speech perception task while keeping both the task and the overall low-level auditory and visual stimuli constant across conditions. We manipulated talker variability by showing participants either the same talker (Same Talker condition) on every trial, or a different talker on every trial (Different Talker condition). Over the entire experiment, each talker was viewed the same number of times in the Same Talker and the Different Talker conditions and thus the individual stimulus properties were held constant. In our experiment, the task was the perception of speech and talker variability was thus an irrelevant factor to the task assigned to the participants. The experiment asked whether manipulating the variability of the talker's identity would induce different viewing strategies.

In order to examine these gaze patterns under varying communication conditions, the intelligibility of the speech was manipulated by adding acoustic noise. When the acoustic speech signal is degraded by the addition of noise, the presence of dynamic visual speech information improves intelligibility (Sumby & Pollack, 1954). The perceptual significance of the visual speech information thus becomes increased when the auditory speech information is degraded (Erber, 1969; O'Neill, 1954, Ross, Saint-Amour, Leavitt, Javitt & Foxe, 2007). Manipulating the intelligibility of speech by the presence of acoustic noise has been shown to alter the spatial distribution and the duration of fixations during audiovisual speech perception tasks (Buchan et al. 2007, Vatikiotis-Bateson et al., 1998). In addition to being assigned to either the Same or Different Talker condition, subjects were also assigned to one of two noise listening conditions (either Noise Absent or Noise Present) in order to determine the effects of both noise and talker variability on behavioral gaze strategies.

The location of gaze fixation is only one indication of what information is being selectively gathered for processing. The duration of fixations may also be partially under cognitive control (Rayner, Liversedge, White & Vergilino-Perez, 2003; Henderson & Pierce, 2008), as opposed to purely stimulus-driven, and the duration of fixations has been shown to vary in audiovisual speech perception under different conditions (Buchan, Paré & Munhall, 2007). To determine whether the gathering of visual information was affected by manipulating talker variability we looked at the overall number of fixations and the duration of fixations in each condition. We also measured the total number and duration of fixations falling in the previously defined regions of interest for each of the features of the face (i.e., the talker's right and left eyes, nose and mouth. See Experimental procedures for details). Additionally, since research suggests that a small number of fixations are purportedly all that is needed to recognize a face (Hsiao & Cottrell, 2007), we measured the number of first and second fixations falling in each region of interest.

2.0 Results

The addition of a multi-talker babble noise reduced the intelligibility of the speech from a mean of 96.8% in the absence of noise to a mean of 40.0% in the presence of noise. Thus performance was significantly lower when acoustic noise was present [$F(1, 124) = 807.057$, $p < .001$]. There was no difference between intelligibility scores for participants who saw the same talker every time as compared with those that saw a different talker every time ($p > .05$), nor was there a significant interaction between the talker variability condition and the presence or absence of noise ($p > .05$).

Manipulating the presence of acoustic noise had a much greater effect on fixations than manipulating talker variability. This is shown descriptively in Figure 1. There are fewer fixations overall when noise is present [F (1,124) = 45.344, p< .001], and the median duration of these fixations is longer than when noise was absent [F (1,124) = 18.284, p< .001]. Note however, that the modal fixation duration remains between 200 and 300 ms for all conditions. As shown in Figure 1, the increase seen in the median duration of fixations is due to a greater number of longer fixations, and not an increase in the peak of the distribution of the durations of the fixations. Despite the large number of subjects in each condition (sixty-four in each of the Same Talker and Different Talker conditions), whether subjects saw the same talker every time, or a different talker every time (i.e. talker variability) had a surprisingly small effect on gaze behavior. There is no significant effect of talker variability on either the overall number of fixations or the median duration of fixations (p> .05). Detailed results for each region of interest (ROI) are presented below for the overall number of fixations and fixation durations, and for the number of first and second fixations in each trial.

In previous studies looking at gaze behavior during face perception, specifically in studies of audiovisual speech perception (Everdell, Marsh, Yurick, Munhall & Paré, 2007; Vatikiotis-Bateson et al., 1998; Paré et al., 2003) and identity judgment (Henderson et al., 2005; Barton et al., 2006; Althoff & Cohen, 1999), there is a preference to fixate the talker's right eye more often than the left eye. We wanted to see if this preference to fixate the right eye also occurred in the current study. In order to do this we computed an asymmetry index (for details see Experimental procedure 4.6.1) for each participant (Everdell et al., 2007) to determine whether a greater proportion of fixations fall on the left or right eye. Since noise produces dramatic reductions in fixations on the eyes (Buchan et al., 2007), an asymmetry index (Everdell et al., 2007) was computed for subjects who were in the Noise Absent condition only. Three subjects (one in the Different Talker condition and two in the Same Talker condition) did not make any fixations on the eyes and were thus excluded from this analysis. Of the remaining subjects in the noise absent condition, seventy percent of the subjects in the same-talker condition and seventy-one percent of subjects in the different-talker condition showed a preference for the right eye, with the remaining subjects showing a preference for the left eye. There was no significant difference between the talker variability conditions [t (59)= 0.322, p=.748]. Using the pooled Same and Different Talker asymmetry indices we found that gaze was significantly biased to the talker's right eye [t (60)= 4.321, p<.001]. Because of this bias, fixations for the right and left eyes were analyzed separately in the region of interest analyses.

2.1 Region of interest Fixations Results

The presence of noise had a noticeable effect on both the number and median duration of fixations in the individual regions of interest. In the absence of noise, there were more fixations on each eye and the mouth, and fewer fixations on the nose than in the presence of noise (see Figure2A). In the presence of noise, fixations on each eye were shorter, and fixations on the nose and mouth were longer (see Figure 3A).

Specifically, when acoustic noise was present there were fewer fixations on the talker's right eye [F (1, 124) = 44.512, p< .001] and left eye [F (1, 124) = 28.287, p< .001]. There were also fewer fixations on the mouth [F (1,124) = 27.013, p< .001] in the presence of noise. Fixations on the right [F (1, 124) = 4.036, p= .047] and left [F (1, 124) = 10.744, p= .001] eye were also shorter when noise was present. The opposite pattern is seen with fixations on the nose (see Figure 4A). In the presence of noise there were a greater number fixations [F (1, 124) = 4.464, p= .037] on the nose, and as can be seen in Figure 4A, the median fixation duration of these fixations on the nose was over twice as long compared with fixations when noise was absent [F (1, 124) = 17.541, p< .001]. For the mouth, there are fewer fixations in the presence of noise

[F (1, 124) = 27.013, p< .001], but these fixations were much longer when noise was present than when it was absent [F (1, 124) = 49.737, p< .001].

There is an overall tendency to fixate more on the features of the face when the talker is varied versus held constant on each trial (see Figure 2B). However, this tendency is only significant for fixations on the mouth [F (1, 124) = 4.216, p= .042]. There was no effect of talker variability on the number of fixations on either the eyes or the nose. Fixations on the mouth were influenced both by talker variability and by the presence of acoustic noise. Additionally, there was also a significant interaction between these two factors [F (1,124) = 4.216, p= .042] (see Figure 2C). When noise was absent, but not when noise was present, there were more fixations on the mouth when subjects saw a different talker every time than when they saw the same talker every time. There was no effect of talker variability on the duration of fixations in any of the regions of interest (see Figure 3B).

2.2 Number of First and Second Fixations in Each Trial by Region of Interest

The presence of noise resulted in a significant decrease in the number of first fixations falling on the right eye [F (1, 124) = 24.662, p< .001 and the left eye [F (1, 124) = 13.555, p< .001]. The same decrease in fixations on the eyes is seen for the number of second fixations falling on the right eye [F (1, 124) = 16.812, p< .001] and the left eye [F (1, 124) = 23.100, p<. 001]. There was an increase in the number of fixations falling on the nose for both the first [F (1, 124) = 45.675, p< .001] and second fixations [F (1, 124) = 23.100, p<. 001]. On the mouth, there was also a significant decrease in the number of second fixations [F (1, 124) = 8.165, p= .005], but not first fixations. There was a modest significant interaction of talker variability with noise on the number of second fixations falling on the mouth region of interest [F 1,124) = 4.109, p=.045] (See Figure 5C). When noise was absent, more of the second fixations in each trial fell on the mouth when subjects saw a different talker every time than when they saw the same talker every time. As with the overall fixations, the presence of noise had a significant effect on the number of first and second fixations falling in each region of interest (see Figures 4A and 5A). The pattern is similar to that seen for the overall fixations

In spite of the fact that the first and second fixations on a face have been shown to play a role in static face processing and recognition, changing the identity of the talker on every trial did not significantly (p>.05) affect how the visual information was gathered during the first and second fixations (see Figures 4B and 5B).

3.0 Discussion

Both varying talker identity and altering the intelligibility of speech by the addition of acoustic noise had an effect on gaze behavior. The presence of noise had a rather dramatic effect on both the number of fixations and the median durations of those fixations. When noise was present participants generally avoided fixating on the eyes as shown by fewer, and shorter, fixations in that region. Participants also made more fixations on the center part of the face in the presence of acoustic noise as indicated by an increase in the number of fixations on the nose. Fixations made on the nose and mouth were also longer in the presence of noise than in the absence of noise. Thus, when intelligibility of the speech was decreased by the addition of the multi-talker babble, subjects adopted a vantage point that was more centralized on the face by reducing the frequency of the fixations on the eyes and mouth and lengthening the duration of their gaze fixations on the nose and mouth. By contrast, varying the identity of the talker had a more modest influence on gaze fixations, and these effects were moderated by the presence of noise. Overall there was a slight increase in the number of fixations on the mouth when viewing a different talker on every trial, particularly in the absence of noise. Surprisingly, there was no effect of talker variability on the first fixation in each trial, and the second fixations showed no main effect and only a small interaction effect with noise level.

Despite the difference in gaze fixations between the two talker variability conditions, there was no difference in intelligibility scores between subjects who saw the same talker every trial versus a different talker every trial. Familiarity with a talker's voice, can help us better understand speech that has been degraded by the presence of acoustic noise (Nygaard & Pisoni, 1998; Nygaard, Sommers & Pisoni, 1994). However, the exposure to the talker in the Same Talker condition in this study was presumably too brief to produce significant speech-related learning. It is possible that the subtle change in visual information gathering strategy when presented with a different talker on every trial was due to increased effort by subjects to gather visual speech and identity information to integrate with the auditory speech and identity information.

Research on facial recognition and face learning shows that the eyes contain a great deal of diagnostic information for making identity judgments (Schyns, Bonnar & Gosselin, 2002; Vinette, Gosselin & Schyns, 2004). The eyes are also preferentially fixated during facial recognition (Henderson, Williams & Falk, 2005; Barton et al. 2006; Althoff & Cohen, 1999) and face learning tasks (Henderson, Williams & Falk, 2005). In our study, however, varying the identity of the talker did not result in differential fixations on the eyes. The only significant effect of manipulating identity was on fixations on the mouth. One possible explanation for this is that facial identity is contained not only in the form of the face, but also in the motion of the face (Hill & Johnson, 2001; Knappmeyer, Thornton & Bülthoff, 2003). In a talking face, most of the motion is in the lower half of the face from the mouth and jaw. The lower part of the face is the major source of visual speech information, with lip movements providing the strongest correlation with the acoustics (Yehia, Rubin & Vatikiotis-Bateson, 1998). To date, the active visual exploration of dynamic faces has only been investigated during audiovisual and visual-only speech perception (e.g. Vatikiotis-Bateson et al., 1998; Lansing & McConkie, 1999, 2003; Paré et al., 2003; Buchan et al., 2007; Everdell et al., 2007) and so it is still unknown how an explicit identity task using dynamic facial stimuli might influence gaze behavior.

A second, related possibility for the difference in mouth fixations is that the identity of a talking face is not restricted to the visual modality. Voices also contain information about identity and gaze fixations may be tuned to maximize the audiovisual integration of this information. The size and shape of a person's vocal tract and vocal cords determine the particular resonance pattern of the voice. Even the shape of a person's teeth and tongue will have an effect on the speech sounds that are produced. Familiarity, and thus identity, can also affect how audiovisual speech is processed. A familiar face matched with an unfamiliar voice is less effective at eliciting the McGurk effect than an unfamiliar face paired with an unfamiliar voice (Walker, Bruce & O'Malley, 1995), suggesting that facial identity and vocal identity are not processed completely independently of one another. A person's facial identity and vocal identity during speech share the same general dynamic temporal patterns. Videos of a person's face talking without sounds can be reliably matched above chance to audio of a person's speech, and vice versa (Lachs & Pisoni, 2004). This cross-modal identity matching of speech can also occur when different utterances are used for each modality (Kamachi et al., 2003). Additionally, familiarity with lip-reading a talker can later help subjects to understand the same talker's auditory speech when their speech has been degraded by acoustic noise (Rosenblum, Miller & Sanchez, 2007). The increased number of mouth fixations when the talker is varied may be related to this process of integrating auditory and visual speech information.

Although changing identity did not have a dramatic effect on gaze fixations, the effect of decreasing the intelligibility of speech by the addition of acoustic noise was quite marked. Our findings in this study are consistent with other eyetracking studies that have looked at the effect of decreasing intelligibility by the addition of acoustic noise. In a study using extended monologues, Vatikiotis-Bateson et al. (1998) saw a decreased number of transitions between areas of the face. Although they did not directly measure the number and duration of fixations,

this is nonetheless consistent with the substantially increased median fixation durations on both the nose and mouth found in the current study. Buchan et al. (2007) used emotionally expressive talking faces and also saw an increase in median fixation duration on the nose and mouth when noise was added.

The increase in gaze fixations on the central part of the face is consistent with the fact that direct foveation of the mouth is not required in order to gather visual speech information (Paré et al., 2003). Rather crude video of the face, which has been degraded by either pixilation (MacDonald, Andersen & Bachmann, 2000) or by spatial frequency filtering (Munhall, Kroos, Jozan & Vatikiotis-Bateson, 2004), provides sufficient visual speech information to influence the perception of speech. Although in our experiment there are fewer fixations made on the mouth when speech was presented in noise, fixations that do fall on the mouth are considerably longer. Additionally there are more and longer fixations on the nose in the presence of noise, and dramatically fewer fixations on the eyes. Taken together, this suggests a strategy where the central and lower parts of the face are being monitored preferentially. Visual speech information is broadly distributed across the face (Yehia et al. 1998), although the lower half of the face contains the lion's share of the information (Thomas & Jordan, 2004).

The shift to monitoring the lower part of the face suggests a subtle shift in strategy where subjects may be preferentially gathering information that provides a benefit to intelligibility. The fact that gaze behavior isn't entirely shifted to the mouth, but rather seems to be clustered around the nose suggests that subjects might be using the nose as a central vantage point that permits a monitoring the eyes and the face for social information while still moving somewhat closer to the lower part of the face. The shift to more centralized fixations on the face is also associated with an overall increase in the duration of fixations. As previously mentioned, this increase in the duration of fixations is due to a greater number of longer fixations, and not to a change in the peak of the distribution of the durations of fixations. The peak of the distribution of the duration of fixations remains between 200 and 300 ms across both talker variability and noise listening conditions. It is possible that this distribution of fixation durations reflects longer fixations that are under direct cognitive control and shorter fixations that are generated through some more automatic mechanism (see Henderson & Pierce, 2008).

Faces contain a wealth of information about speech, emotion, and gaze direction in addition to information about identity. Anatomy constrains how the face can vary to display this information, and so different forms of social information contained in the face must often spatially and temporally coexist. Visual information gathering strategies likely optimize the collection of all information available on the face, balancing the gathering of task-relevant information with extracting maximal information for social and communication interactions. Such a strategy would involve the systematic scanning of facial regions as well as parallel processing of foveal and peripheral visual cues. While the extraction of task relevant information biases gaze behavior, the rather modest effect of task instructions hints at a single strategy to extract as much facial information as possible.

Studying identity using static images has yielded a great number of insights, yet static stimuli are nonetheless impoverished versions of the normally dynamic face, lacking much of the information available in the dynamic face. Because of this, behavioral strategies to gather information from static faces may not be reflected in how visual information is gathered from dynamic faces.

4.0 Experimental Procedure

4.1 Subjects

One hundred and twenty eight individuals (86 females) with a mean age of 20.8 years of age participated in this experiment. All subjects were native speakers of English and reported having normal or corrected to normal vision, as well as no speech or hearing difficulties. The experiment was undertaken with the understanding and written consent from each subject. Subjects were fully informed about the experimental procedures, but were only informed about the specific hypothesis after they had completed the study.

4.2 Stimuli

The stimuli consisted of sixteen talkers (8 female) who were filmed in color using digital audio and video recording equipment. Talkers said the low context sentences drawn from form 2.5 of the Speech in Noise (or SPIN) sentences (Kalikow, Stevens & Elliott, 1977). The number of last words correct was used as a measure of intelligibility. For example, in the sentence “Miss White would consider the mold”, only the word mold was scored. For half of the participants, the intelligibility of the speech was reduced by the addition of a multi-talker noise babble signal (ten male and ten female talkers) (Auditec, St. Louis, MO). The presentation level for the noise absent condition was 61 dB(A), and the presentation level for the noise present condition was 70 dB(A).

4.3 Experimental task

The experiment took place in a double-walled sound booth. Subjects were seated approximately 57 cm away from the centre of a 20-in television monitor (Sony PVM 20L5). Subjects' heads were stabilized with a chinrest. The audio signal was played from speakers (Paradigm Reference Studio/20) positioned on either side of the monitor. Eye position was monitored using an Eyelink II eye-tracking system (SR Research, Osgoode, Canada) using dark pupil tracking with a sampling rate of 500Hz. A nine-point calibration and validation procedure was used. The maximum error on a single point was 1.5 visual degrees, though the error on the central point was always less than 1 degree. A drift correction was performed before each trial.

The experiment was carried out as a between-subjects design with subjects assigned to one of two talker variability conditions. Subjects either saw the same talker on each trial (Same Talker condition), or a different talker each trial (Different Talker condition). There were sixteen different talkers who each said the same sixteen sentences. Each trial consisted of a sentence, and each sentence was used only once per subject. A between-subjects design was used as this avoided having subjects' knowledge of the other conditions biasing their gaze behavior. Subjects were told that they would see someone on the screen saying a sentence on each trial. Subjects were instructed to watch the talker the entire time they were on the screen, and to report all of the words that they heard them say.

4.4 Scoring of speech task

As is standard for the SPIN sentences only accuracy of the perception of the last word were used as a measure of intelligibility. Subject responses on the speech task were analyzed using a 2X2 (Talker variability condition X noise listening condition) ANOVA.

4.5 Coding of facial features and regions of interest

Instantaneous positions of the right and left eyes, the nose and the mouth were coded frame-by-frame for each sentence of the stimuli. One reference point was coded for each eye, approximately in the centre of the pupil. For the nose, a point was coded for the outside of the

widest part of each nostril, and a virtual point approximately 0.8 degrees of visual angle above the halfway point was chosen to represent the nose feature. For the mouth, four points were coded, one point in each of the corners of the mouth, one on midline of the upper lip on the vermillion border, and one on the midline of the lower lip on the vermillion border.

Ellipses centered on each eye, the nose, and the mouth reference points were used to delimit regions of interest. For each eye, an ellipse with a vertical semi-minor axis corresponding to approximately 2.0 degrees of visual angle, and a horizontal semi-major axis of 1.4 visual degrees was used to demarcate the region of interest boundary. The ellipse centered on the nose reference point also had a vertical semi-minor axis of 2.0 degrees of visual angle, and a horizontal semi-major axis of 1.4 degrees of visual angle. Because the mouth can change shape quite considerably during speech, the mouth region of interest was variable in size. A centre point, determined by the position of the four points that had been coded on the mouth, was used to centre an ellipse that was 0.8 degrees of visual angle larger from the centre point than each of the four coded points. See Figure 6 for an illustration of the regions of interest.

A further region of interest, encompassing the face, was created for quality control purposes. For the face, an ellipse was centered on the nose, with a horizontal semi-major axis of 5.6 visual degrees corresponding to approximately 5.6 degrees of visual angle, and a vertical semi-major axis corresponding to approximately 8.5 degrees of visual angle. For the duration of the trials during the experiment a mean of 99% (range 87-100%) of eyetracker samples recorded from each participant from fell on the face region of interest.

4.6 Dependant measures of gaze behavior

4.6.1. Asymmetry index—The asymmetry index is a ratio of the number of fixations on the right and left eyes: index= (right eye - left eye)/(left eye- right eye). Positive values show a bias towards the right eye and negative values towards the left eye. The effect of talker variability was analyzed with a t-test. The significance of the bias of the asymmetry index was computed with a one-sample test with a test value of 0.

4.6.2 Fixation analyses—Fixations were analyzed over the entire trial. The mean number and median duration of gaze fixations falling in each region of interest were calculated as well as the mean distance of gaze fixations with respect to the centre of each region of interest. The effect of talker variability on the above measures was analyzed with 2X2 (talker variability condition X noise listening condition) ANOVAs. Each region of interest was analyzed independently.

4.6.3 First and second Fixation analyses—In a separate analysis, we measured the number of first and second fixations in each trial that fell in each region of interest. The effect of talker variability and noise level was analyzed with 2X2 (talker variability condition X noise listening condition) ANOVAs. Each region of interest was analyzed independently.

Acknowledgements

The National Institute on Deafness and other Communication Disorders (grant DC-00594), the Natural Sciences and Engineering Research Council of Canada, and the Canadian Institutes of Health Research supported this work.

Thank you to Katrine Doucet for her help in creating the stimuli. Thank you to Dave Hoffmann and Fred Kroon for their help with data analysis. M.P. holds a New Investigator Award from the Canadian Institutes of Health Research. J.B. held an Ontario Graduate Scholarship for part of the duration of the study and currently holds a Natural Sciences and Engineering Research Council of Canada PGS D3 award and the Brian R. Shelton Graduate Fellowship.

References

Althoff RR, Cohen NJ. Eye-movement-based memory effect: a reprocessing effect in face perception. J Exp Psychol Learn 1999;25:997–1010.

Ambadar Z, Schooler JW, Cohn JF. Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions. Psychol Sci 2005;16:403–410. [PubMed: 15869701]

Baron-Cohen S, Wheelwright S, Hill J, Raste Y, Plumb I. The "Reading the mind in the eyes" test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. J Child Psychol Psychiatr 2001;42:241–251.

Barton J, Radcliffe N, Cherkasova M, Edelman J, Intriligator J. Information processing during face recognition: the effects of familiarity, inversion and morphing on scanning fixations. Perception 2006;353:1089–1105. [PubMed: 17076068]

Benoît, C.; Guiard-Marigny, T.; Le Groff, B.; Adjoudani, A. Which components of the face do humans and machines best speechread?. In: Stork, DG.; Hennecke, M., editors. Speechreading by Humans and Machines: Models, Systems & Applications. Kluwer Academic Publishers; 1996.

Bernstein LE, Demorest ME, Tucker PE. Speech perception without hearing. Percept Psychophys 2000;62:233–252. [PubMed: 10723205]

Buchan JN, Paré M, Munhall KG. Spatial statistics of gaze fixations during dynamic face processing. Soc Neurosci 2007;2:1–13. [PubMed: 18633803]

Emery NJ. The eyes have it: the neuroethology, function and evolution of social gaze. Neuroscience Behav R 2000;24:581–604.

Erber NP. Interaction of audition and vision in the recognition of oral speech stimuli. J Speech Hear Res 1969;12:423–425. [PubMed: 5808871]

Everdell IT, Marsh H, Yurick M, Munhall KG, Paré M. Gaze behavior in audiovisual speech perception: Asymmetrical distribution of face-directed fixations. Perception 2007;36:1535–1545. [PubMed: 18265836]

Findlay, JM.; Gilchrist, ID. Active vision: The psychology of looking and seeing. Oxford University Press; London: 2003.

Ghazanfar AA, Turesson HK, Maier JX, van Dinther R, Patterson RD, Logothetis NK. Vocal-tract resonances as indexical cues in rhesus monkeys. Curr Biol 2007;17:425–430. [PubMed: 17320389]

Henderson JM. Human gaze control during real-world scene perception. Trends Cogn Sci 2003;7:498–504. [PubMed: 14585447]

Henderson JM, Pierce GL. Eye movements during scene viewing: Evidence for mixed control of fixation durations. Psychon B Rev 2008;15:566–573.

Henderson JM, Weeks PA Jr, Hollingworth A. The effects of semantic consistency on eye movements during scene viewing. J Exp Psychol Human 1999;25:210–228.

Henderson JM, Williams CC, Falk RJ. Eye movements are functional during face learning. Mem Cognition 2005;33:98–106.

Hill H, Johnson A. Categorizing sex and identity from the biological motion of faces. Curr Biol 2001;11:880–885. [PubMed: 11516651]

Hsiao JH, Cottrell GW. The influence of number of eye fixations on face recognition. J Vision 2007;7:494a.

Lachs L, Pisoni DB. Cross-modal source information and spoken word recognition. J Exp Psychol Human 2004;30:378–396.

Lander K, Bruce V. Recognizing famous faces: Exploring the benefits of facial motion. Ecol Psychol 2000;14:385–388.

Lansing CR, McConkie GW. Attention to facial regions in segmental and prosodic visual speech perception tasks. J Speech Lang Hear R 1999;42:526–539.

Lansing CR, McConkie GW. Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. Percept Psychophys 2003;65:536–552. [PubMed: 12812277]

Kalikow DN, Stevens KN, Elliott LL. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. J Acoust Soc Am 1977;61:1337–1351. [PubMed: 881487]

Kamachi M, Hill H, Lander L, Vatikiotis-Bateson E. 'Putting the face to the voice': Matching identity across modality. Curr Biol 2003;13:1709–1714. [PubMed: 14521837]

Knappmeyer B, Thornton IM, Bülthoff HH. The use of facial motion and facial form during the processing of identity. Vision Res 2003;43:1921–1936. [PubMed: 12831755]

MacDonald J, Andersen S, Bachmann T. Hearing by eye:how much spatial degradation can be tolerated? Perception 2000;29:1155–1168. [PubMed: 11220208]

MacDonald J, McGurk H. Visual influences on speech perception processes. Percept Psychophys 1978;24:253–257. [PubMed: 704285]

McGurk H, MacDonald JW. Hearing lips and seeing voices. Nature 1976;264:746–748. [PubMed: 1012311]

Munhall KG, Kroos C, Jozan G, Vatikiotis-Bateson E. Spatial frequency requirements of audiovisual speech perception. Percept Psychophys 2004;66:574–538. [PubMed: 15311657]

Nygaard LC, Pisoni DB. Talker-specific learning in speech perception. Percept Psychophys 1998;60:355–376. [PubMed: 9599989]

Nygaard LC, Sommers MS, Pisoni DB. Speech perception as a talker-contingent process. Psychol Sci 1994;5:42–46.

O'Neill JJ. Contributions of the visual components of oral symbols to speech comprehension. J Speech Hearing Disord 1954;19:429–439. [PubMed: 13222457]

O'Toole AJ, Roark DA, Abdi H. Recognizing moving faces: A psychological and neural synthesis. Trends Cogn Sci 2002;6:261–266. [PubMed: 12039608]

Paré M, Richler RC, ten Hove M, Munhall KG. Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. Percept Psychophys 2003;65:533–567.

Parkhurst D, Law K, Niebur E. Modeling the role of salience in the allocation of overt visual attention. Vision Res 2002;42:107–123. [PubMed: 11804636]

Parkhurst DJ, Niebur E. Scene content selected by active vision. Spatial Vision 2003;16:125–154. [PubMed: 12696858]

Rayner K, Liversedge SP, White SJ, Vergilino-Perez D. Reading disappearing text: cognitive control of eye movements. Psychol Sci 2003;14:385–388. [PubMed: 12807416]

Rosenblum LD, Miller RM, Sanchez K. Lip-read me now, hear me better later: cross-modal transfer of talker-familiarity effects. Psychol Sci 2007;18:392–396. [PubMed: 17576277]

Thomas SM, Jordan TR. Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. J Exp Psychol Human 2004;30:873–888.

Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, Foxe JJ. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cereb Cortex 2007;17:1147–1153. [PubMed: 16785256]

Schyns PG, Bonnar L, Gosselin F. Show me the features! Understanding recognition from the use of visual information. Psychol Sci 2002;13:402–409. [PubMed: 12219805]

Schweinberger SR, Soukup GR. Asymmetric relationships among perceptions of facial identity, emotion and facial speech. J Exp Psychol Human 1998;24:1748–1795.

Stacey PC, Walker S, Underwood JD. Face processing and familiarity: evidence from eye-movement data. Br J Psychol 2005;96:407–422. [PubMed: 16248933]

Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. J Acoust Soc Am 1954;26:212–215.

Vatikiotis-Bateson E, Eigsti IM, Yano S, Munhall KG. Eye movement of perceivers during audiovisual speech perception. Percept Psychophys 1998;60:926–940. [PubMed: 9718953]

Vinette C, Gosselin F, Schyns PG. Spatio-temporal dynamics of face recognition in a flash: it's in the eyes. Cognitive Sci 2004;28:289–301.

Walker S, Bruce V, O'Malley C. Facial identity and facial speech processing: familiar faces and voices in the McGurk effect. Percept Psychophys 1995;57:124–1133.

Walker-Smith GJ, Gale AG, Findlay JM. Eye movement strategies involved in face perception. Perception 1977;6:313–326. [PubMed: 866088]

Yakel DA, Rosenblum LD, Fortier MA. Effects of talker variability on speechreading. Percept Psychophys 2000;62:1405–1412. [PubMed: 11143452]

Yarbus, AL. Eye movements and vision. New York: Plenum Press; 1967.

Yehia HC, Rubin PE, Vatikiotis-Bateson E. Quantitative association of vocal-tract and facial behavior. Speech Commun 1998;26:23–44.

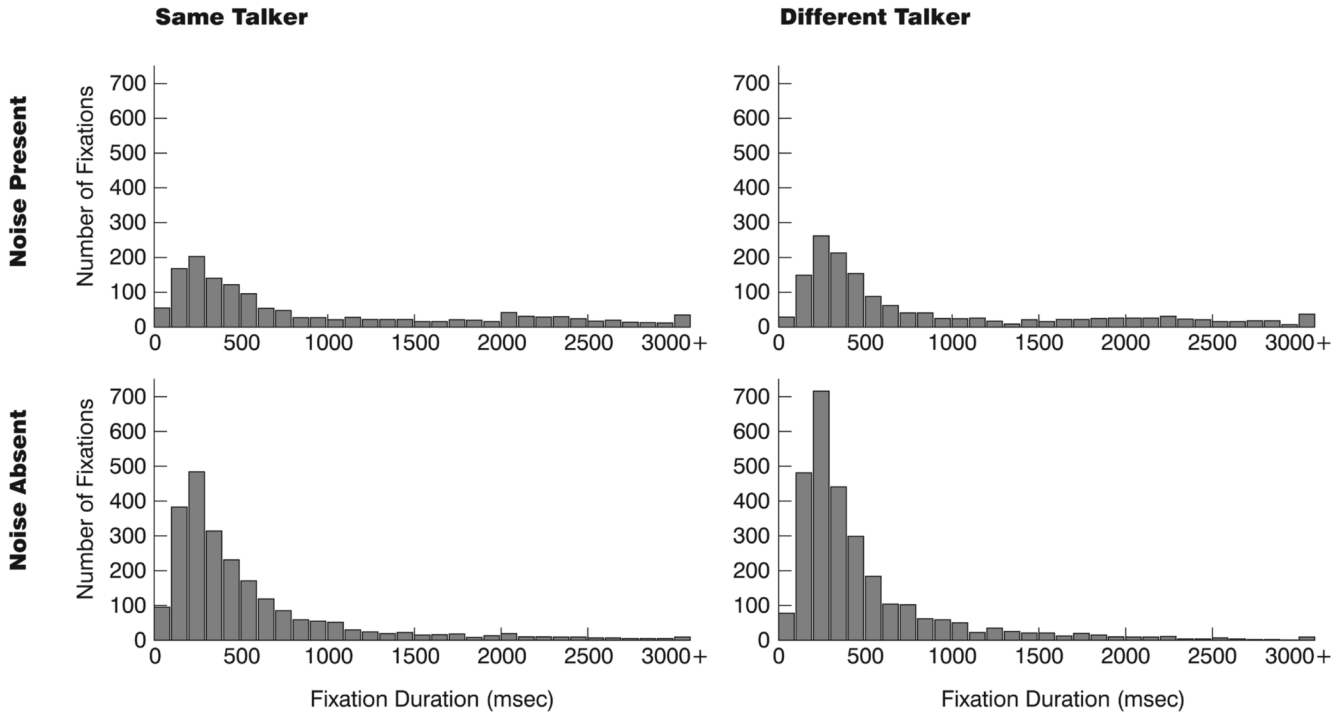

Figure 1. Histogram of all fixations by experimental condition

The distribution of fixation durations for each experimental condition is shown.

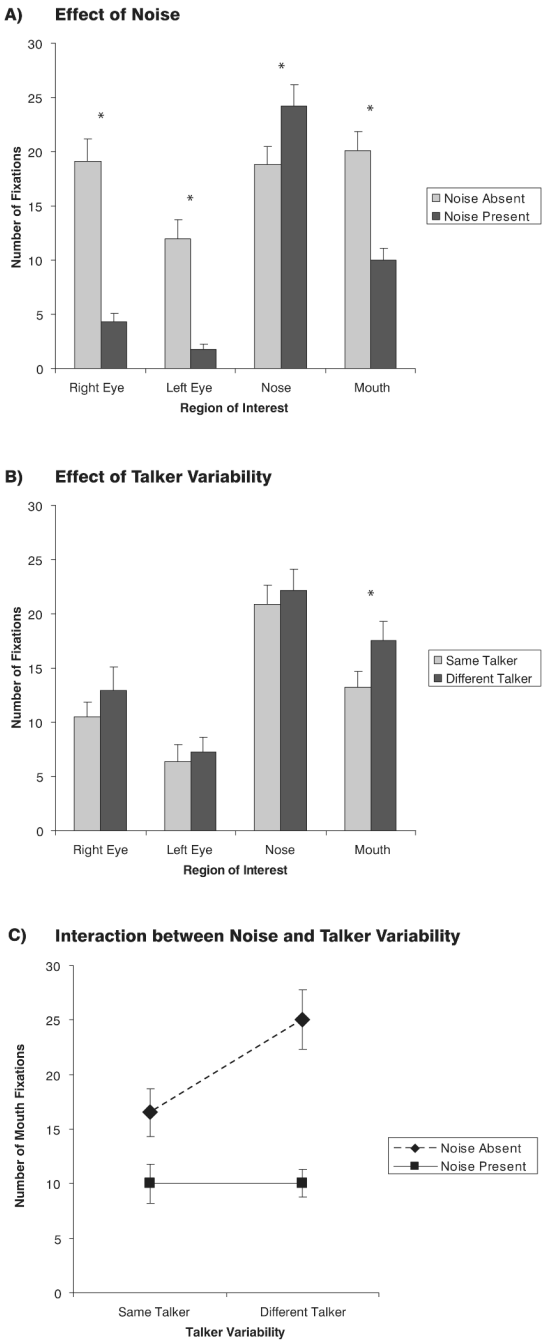

Figure 2. Overall number of fixations by region of interest

The overall number of fixations in each region of interest is shown. 2A shows gaze as a function of noise and 2B shows gaze as a function of talker variability. 2C shows the significant interaction between noise and talker variability on the number of gaze fixations made in the mouth region of interest. An asterisk denotes significant differences. The error bars indicate the standard errors of the mean.

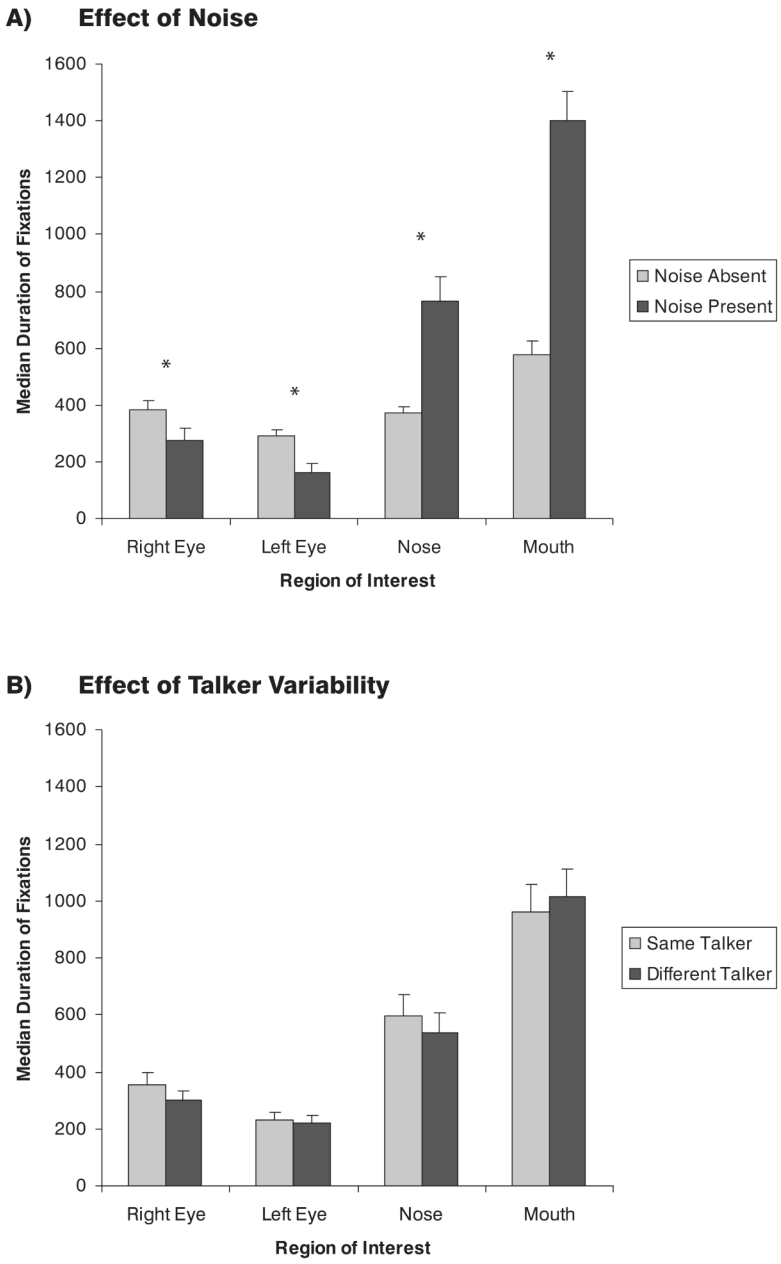

Figure 3. Overall duration of fixations by region of interest

The median durations of fixations falling in each region of interest is shown. 3A shows gaze as a function of noise and 3B shows gaze as a function of talker variability. An asterisk denotes significant differences. The error bars indicate the standard errors of the mean.

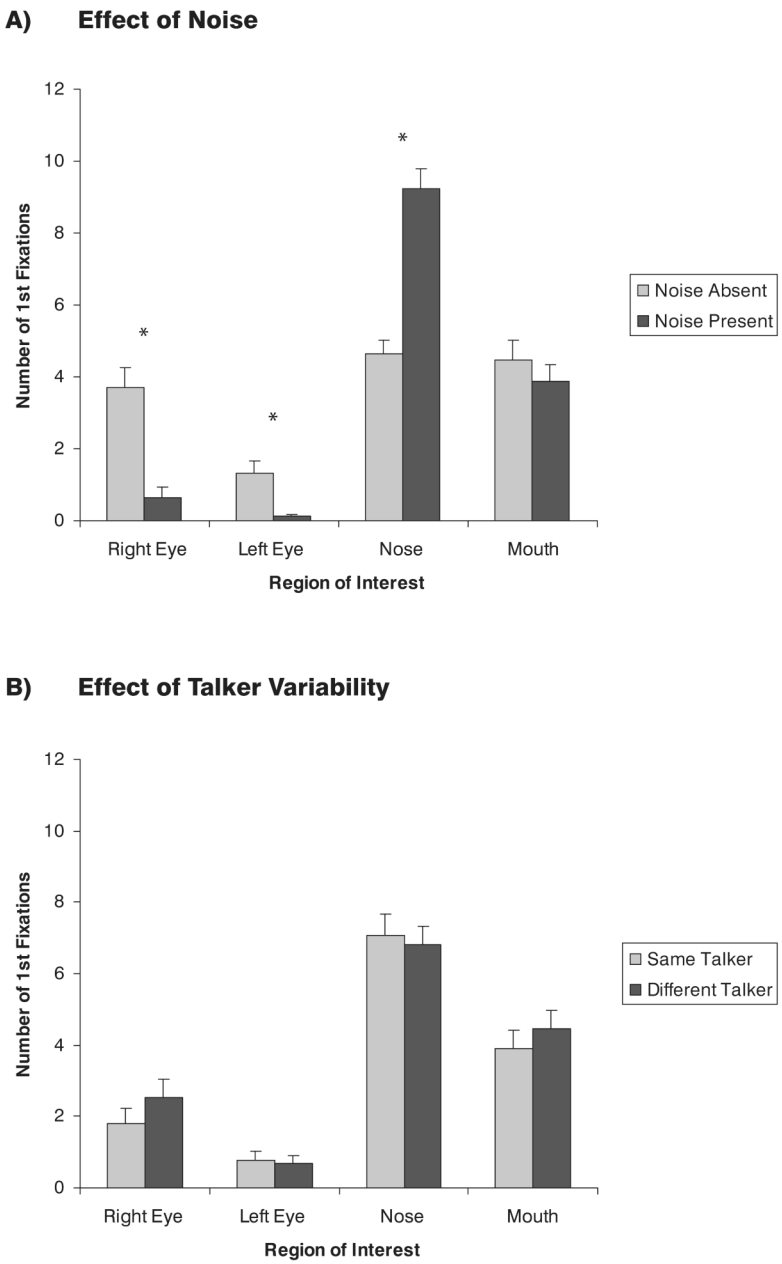

Figure 4. Number of first fixations by region of interest

The number of first fixations in each trial falling in each region of interest is shown. 4A shows gaze as a function of noise and 4B shows gaze as a function of talker variability. An asterisk denotes significant differences. The error bars indicate the standard errors of the mean.

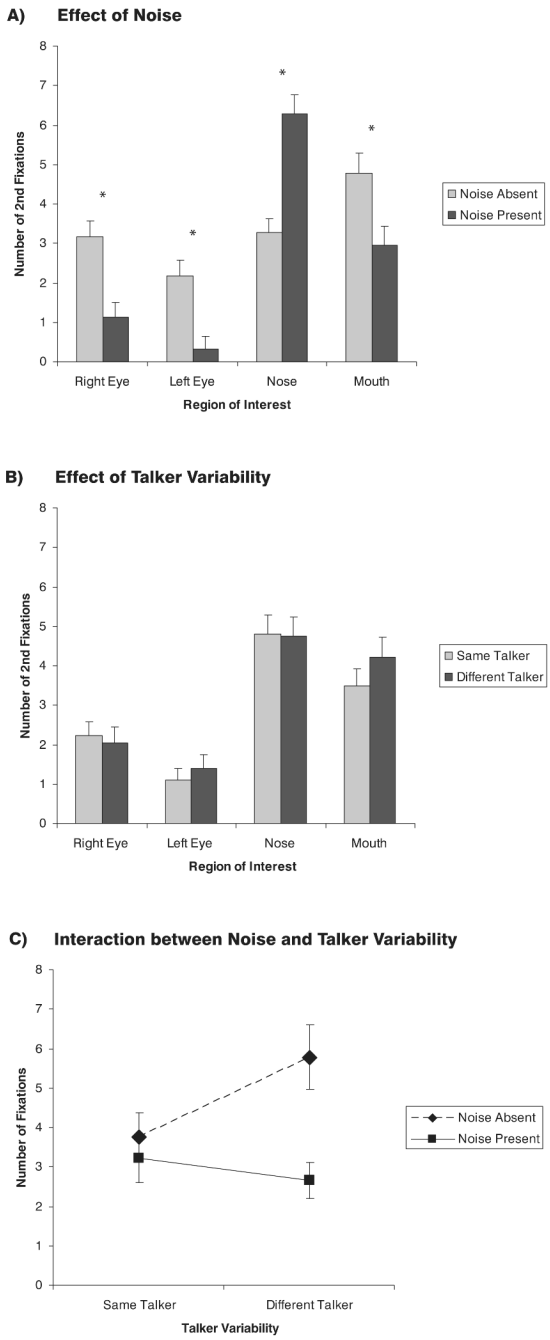

Figure 5. Number of second fixations by region of interest

The number of second fixations in each trial falling in each region of interest is shown. 5A shows gaze as a function of noise and 5B shows gaze as a function of talker variability. 5C shows the significant interaction between noise and talker variability on the number of gaze fixations made in the mouth region of interest. An asterisk denotes significant differences. The error bars indicate the standard errors of the mean.

Figure 6. Regions of interest

The ellipses indicate the regions of interest (the talker's right eye and left eye, nose and mouth) used in the analyses for four of the sixteen talkers are shown. Videos of the talkers were presented in color. See experimental procedures section for details.