Available online at www.sciencedirect.com
ScienceDirect
Cognition 101 (2006) 515–529
COGNITION
www.elsevier.com/locate/COGNIT

Eye movements of monkey observers viewing vocalizing conspecifics

Asif A. Ghazanfar *, Kristina Nielsen, Nikos K. Logothetis

Max Planck Institute for Biological Cybernetics, Spemannstrasse 38, 72076 Tuebingen, Germany

Received 26 March 2005; revised 29 November 2005; accepted 1 December 2005

Abstract

Primates, including humans, communicate using facial expressions, vocalizations and often a combination of the two modalities. For humans, such bimodal integration is best exemplified by speech-reading – humans readily use facial cues to enhance speech comprehension, particularly in noisy environments. Studies of the eye movement patterns of human speech-readers have revealed, unexpectedly, that they predominantly fixate on the eye region of the face as opposed to the mouth. Here, we tested the evolutionary basis for such a behavioral strategy by examining the eye movements of rhesus monkeys observers as they viewed vocalizing conspecifics. Under a variety of listening conditions, we found that rhesus monkeys predominantly focused on the eye region versus the mouth and that fixations on the mouth were tightly correlated with the onset of mouth movements. These eye movement patterns of rhesus monkeys are strikingly similar to those reported for humans observing the visual components of speech. The data therefore suggest that the sensorimotor strategies underlying bimodal speech perception may have a homologous counterpart in a closely related primate ancestor.

Keywords: Bimodal; Multisensory; Scanpath; Lip-reading; Superior temporal sulcus; Auditory cortex; Frontal eye fields; Crossmodal; Evolution of speech; Speech reading; Primate cognition; Eye gaze

* Corresponding author. Present address: Program in Neuroscience, Department of Psychology, Princeton University, Green Hall, Princeton, NJ 08544, USA. Tel.: +1 609 258 9314.

E-mail address: asifg@princeton.edu (A.A. Ghazanfar).

doi:10.1016/j.cognition.2005.12.007

1. Introduction

For both human and nonhuman primates, everyday social interactions often occur in noisy auditory environments in which the vocalizations of other conspecifics, heterospecifics, abiotic noise, and physical obstructions can degrade the quality of auditory information. This ambient noise presents a serious obstacle to communication in all the natural habitats of primates (Brown, 2003). The auditory perceptual system, consequently, has evolved noise tolerant strategies to overcome these problems. For example, primates can recognize severely degraded vocal signals using temporal cues (Ghazanfar, Smith-Rohrberg, Pollen, & Hauser, 2002; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). Another perceptual mechanism that evolved to compensate for noisy auditory environments is the audiovisual integration of vocal signals. Bimodal vocal signals can offer robust advantages in detection, discrimination and learning, as has been shown for multimodal signals in other domains, modalities, and taxonomic groups (Rowe, 1999).

Watching a speaker's face can enhance perception of auditory speech under ideal (Reisberg, McLean, & Goldfield, 1987) and compromised (Cotton, 1935; Sumby & Pollack, 1954) listening conditions, raising the question of what cues are being used in visual speech perception. One method for investigating the behavioural strategies involved in facial–vocal process is the measurement of eye movement patterns. Recently, studies of human subjects have examined observers' eye movements while viewing talkers in a naturalistic setting (Klin, Jones, Schultz, Volkmar, & Cohen, 2005) or under different listening conditions, including varying levels of background noise (Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998), competing voices (Rudmann, McCarley, & Kramer, 2003), and silence (i.e., speech-reading with no audio track) (Lansing & McConkie, 1999, 2003). When human subjects are given *no* task or instruction regarding what acoustic cues to attend, they will consistently look at the eye region more than the mouth when viewing videos of human speakers (Klin et al., 2005). However, when subjects are required to perform a specific task, then eye movement patterns are task-dependent. For example, when required to attend to speech-specific aspects of the communication signal (e.g., phonetic details in high background noise, word identification or segmental cues), humans will make significantly more fixations on the mouth region than the eye region (Lansing & McConkie, 2003; Vatikiotis-Bateson et al., 1998). In contrast, when subjects are asked to focus on prosodic cues or to make social judgments based on what they see/hear, they direct their gaze more often towards the eyes than the mouth (Buchan, Pare, & Munhall, 2004; Buchan, Pare, & Munhall, 2005; Lansing & McConkie, 1999).

The evolution of sensorimotor mechanisms that analyze and integrate facial and vocal expressions is likely an innovation that is not specific to human speech perception (Ghazanfar & Santos, 2004). Many nonhuman primate species have large and diverse repertoires of vocalizations and facial expressions (Andrew, 1962; Van Hooff, 1962), and often these communication signals are co-occurring (Hauser, Evans, & Marler, 1993; Partan, 2002). The visual and auditory behavior of rhesus monkeys (*Macaca mulatta*), in particular, have been particularly well-studied (Hauser et al., 1993; Hauser & Marler, 1993; Hinde & Rowell, 1962; Partan, 2002; Rowell & Hinde, 1962). As in human speech, when rhesus monkeys produce a particular vocalization, it is often associated with a unique facial posture (Hauser et al., 1993; Partan, 2002). For example, threat calls are accompanied by an open-mouth posture and staring, whereas coo calls are produced with the lips protruded (see Fig. 1A). Furthermore, like human adults and infants (Kuhl, Williams, & Meltzoff, 1991; Patterson & Werker, 2003), rhesus monkeys are able to spontaneously (that is, without any training) match heard vocalizations with the appropriate facial postures (Ghazanfar & Logothetis, 2003). We do not know, however, whether humans and monkeys use the same sensorimotor processes when they view vocalizing conspecifics.

To characterize the similarities and differences between monkey and human audiovisual communication, we investigated the eye movement patterns of rhesus monkeys while they viewed digitized videos of conspecifics producing vocalizations. We generated video sequences of monkeys vocalizing and varied the listening conditions by modifying the audio track. In the first experiment, we varied the background noise levels by mixing in monkey ‘cocktail’ party noise. In the second experiment, we compared responses to normal movie sequences with sequences in which the audio track was either silenced or where the auditory component of the vocalizations were paired with the incorrect facial posture (i.e., mismatched). In both experiments, the monkey subjects were not required to perform a task, but simply free-viewed the videos in whatever spontaneous manner they chose.

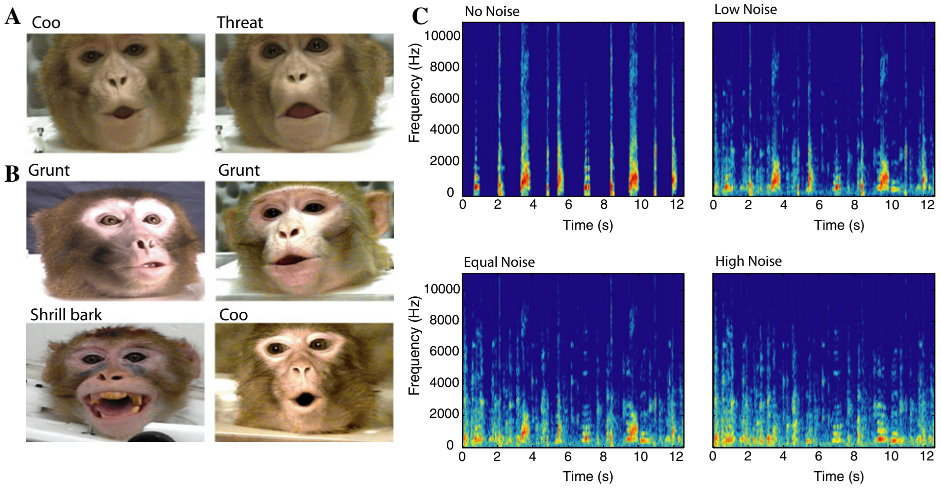

Fig. 1. Stimulus exemplars. (A) Two exemplar expressions from Movie 1, which consisted of the vocalizations of a single monkey. Movie 2, as illustrated by the four examples given in (B), contained video clips of the vocalizations of different individuals. For each movie, four conditions with different levels of background noise were generated. The spectrograms shown in (C) highlight the impact of the background noise on the audio track of Movie 2.

2. Methods

We tested four adult male rhesus macaques (*M. mulatta*) who are part of a large colony housed at the Max Planck Institute for Biological Cybernetics. The eye movements of these four monkeys were recorded with a scleral search coil, which was implanted together with a head-post in a sterile surgery. The subjects all had normal hearing as evidenced by their performance in multiple auditory and auditory-visual behavioural and neurophysiological experiments (Ghazanfar & Logothetis, 2003; Ghazanfar, Maier, Hoffman, & Logothetis, 2005; Ghazanfar, Neuhoff, & Logothetis, 2002; Maier, Neuhoff, Logothetis, & Ghazanfar, 2004). All animals are socially housed and provided with enrichment (toys, hammocks, ropes, etc.). All experimental procedures were in accordance with the local authorities (Regierungspraesidium Tübingen) and the European Community (EUVD 86/609/EEC) for the care and use of laboratory animals.

2.1. Stimuli

The naturalistic stimuli were digital video clips of vocalizations produced by rhesus monkeys in the same colony. Vocalizations included coos, threats, grunts, and a shrill bark. All but one stimulus were filmed with a JVC GR-DVL805 digital camera (www.jvc.com) while monkeys spontaneously vocalized while sitting in a primate restraint chair placed in a sound-attenuated room. This ensured that each video had similar visual and auditory conditions and that the individuals were in similar postures when vocalizing. The one exception was a "shrill bark" call produced by a chair-restrained monkey in a quiet room but with no acoustic treatment. This alarm call is produced infrequently and so was filmed opportunistically. Videos were acquired at 30 frames per second (frame size: 720 × 480 pixels), while the audio tracks were acquired at 32 kHz and 16-bit resolution in mono. Across the vocalizations, the audio tracks were measured and matched in average RMS energy using Adobe Audition 1.0 (www.adobe.com).

Three video sequences were constructed. Movie 1 (10 s) showed vocalizations of a single monkey, while Movie 2 (12.5 s) contained short clips of five different individuals. Species-typical "cocktail party noise" was recorded from our rhesus monkey colony just prior to feeding. During this time, many monkeys will produce food-related calls such as coos and grunts (Hauser & Marler, 1993). We mixed this noise in with the audio track of Movies 1 and 2 at four different relative levels: (1) no noise; (2) low noise – noise level 9 dB below the audio track; (3) equal noise – noise and audio track at equal levels; and (4) high noise – noise track 15 dB higher than the audio track.

Movie 3 was made to test whether the auditory component of the vocalizations has any influence at all on eye movement patterns and to eliminate the potential confound of multiple video edits creating abrupt transitions between vocalizations. A 30-s clip of a single novel vocalizing animal was made using only two edits (i.e., two points where frame transitions are abrupt) during neutral expressions. The vocalizer produced five different calls with large inter-call intervals (see Fig. 6B).

From this video, three audio conditions were tested: (1) Normal (auditory track as originally recorded), (2) Mismatch (call types shuffled but retain the appropriate onset times), and (3) Silence.

2.2. Behavioural apparatus and paradigm

Experiments were conducted in a double-walled sound attenuating booth measuring 1.7 × 2.0 × 2.1 m (l × w × h; inner dimensions), lined with echo-attenuating foam (www.sonex.com). The monkey sat in a primate chair secured in front of a 21-in. color monitor at a distance of 94 cm. Directly on either side of the monitor were two JBL Control 1X speakers (frequency response: 80–20 kHz; www.jbl.com) powered by a Sony Amplifier (TA-FE570; www.sony.com). Two speakers were used to eliminate the spatial mismatch between the visual signals and the auditory signals.

The monkeys performed in a darkened booth. Movies and sound conditions were presented in random order and with a 10 s inter-trial interval. The monkeys did not perform any task, but could freely view the videos with their heads fixed into a forward-facing position. They were not rewarded for their performance, but during the inter-trial interval they were given juice to keep them motivated to stay awake. The videos were displayed with a size of 20 × 13.2 deg and the audio was played at $\sim$72 dB (as measured by a Bruel and Kjaer 2238 Mediator (www.bkhome.com; sound level meter at 94 cm and C-weighted)). We used larger-than-life stimuli to increase the necessity of saccadic eye movements to salient facial features (Vatikiotis-Bateson et al., 1998). Stimuli were presented via a dedicated graphics workstation at a resolution of 1024 × 768 at 75 Hz refresh rate. Behavioral control for the experiments was maintained by a network of inter-connected PCs running the QNX real-time operating system (QSSL, Ontario, Canada). Eye position signals were digitized at a sampling rate of 200 Hz (CNC Engineering, Seattle, WA).

2.3. Data analysis

From the eye movement data, fixation periods were extracted using a velocity criterion. Fixation periods were marked as time periods longer than 100 ms in which no eye movement faster than 20 deg/s occurred; the eye position during a fixation was calculated as the mean eye position during the fixation period. Overlaying the fixation locations on the respective video frames, we counted the number of fixations falling on either the eye or mouth region. In each movie frame, eye and mouth regions were outlined by two ellipses. For the eye region, the ellipse vertices were placed at the outer canthi of the eyes, and the minor axis of the ellipse was defined by the distance from the brow ridge to just above the nostrils. For the mouth region, the major axis fell onto the midline of the mouth, with the vertices at both corners of the mouth, and the minor axis extended from below the nostrils to the chin. These regions were adapted frame-by-frame to account for head movements. When computing the percentage of fixations, we normalized these data by the number of fixations falling onto any point of the image. In addition, we determined the onset of each fixation and the onset of the corresponding mouth movement/vocalization in the video.

To determine whether fixation patterns varied according to call type, we conducted additional analyses on those fixations that fell in the mouth region. Specifically, we examined whether the distribution of mouth fixations varied according to call type and whether this was influenced by background noise. In this analysis, the number of mouth fixations for the different call types were normalized by the total number of mouth fixations during that condition (i.e., noise level) and expressed as a percentage. A second analysis examined whether the duration of fixations that fell upon the mouth varied according to noise levels.

3. Results

For the first experiment, we used two separate movie sequences. In Movie 1 (10 s), the same monkey produces two different coo exemplars and two different threat exemplars (Fig. 1A). This sequence should minimize the patterns of fixation (if any) related to identifying the individual seen by the subject. Movie 2 (12.5 s) consisted of different individuals producing coos, grunts, and shrill barks (Fig. 1B). This movie (presumably) maximized the amount of 'interest' shown by our monkey subjects and thereby reduced the effects of habituation. For each of these two movies, there were four different conditions: (1) 'no noise' condition where only the audio track associated with the faces was heard; (2) 'low noise' where background noise was mixed below the overall level of the original audio track; (3) 'equal noise' whereby the overall levels of both the noise and the original audio track were made to be equivalent; and (4) 'high noise', whereby the noise exceeded the level of the original audio track (Fig. 1C). The 'noise' in these conditions consisted of the natural 'cocktail party' noise of our colony of macaques.

Fig. 2 shows a representative example of the fixation patterns of one monkey subject viewing Movie #1 in the 'equal noise' condition. Only frames in which there was a fixation on some part of the face are shown and each fixation is represented by a single red dot on the monkey's face. Most fixations were on the eye region with occasional fixations on the mouth when the mouth was in motion. This pattern was maintained across all four monkey subjects and for both video sequences (Fig. 3). The fixation patterns did not vary significantly across the different noise conditions: in all conditions, the monkey fixated on the eye region far more than the mouth region. A repeated measures ANOVA with Face Region, Noise Condition, and Movie as factors revealed the main effect of Face Region as significant ($F(1,3) = 98.56, p = 0.002$); neither Noise Condition ($F(1,3) = 0.563, p = 0.653$) nor Movie ($F(1,3) = 1.77, p = 0.276$) had significant main effects. Furthermore, there were no significant interactions (all p values between 0.28 and 0.95). Thus, for all monkeys and under all conditions, the eyes were fixated proportionally more than the mouth during viewing of vocalizing conspecifics.

We measured the time into the trial when the monkeys made their first fixations on the eyes and the mouth region (Fig. 4). Monkeys rapidly fixated the eye region, on average 0.42 s into the trial. In contrast, the mouth region was fixated upon with a mean of 4.17 s into the trial. This time difference was highly significant ($t(56) = 8.563$, $p < 0.000$).

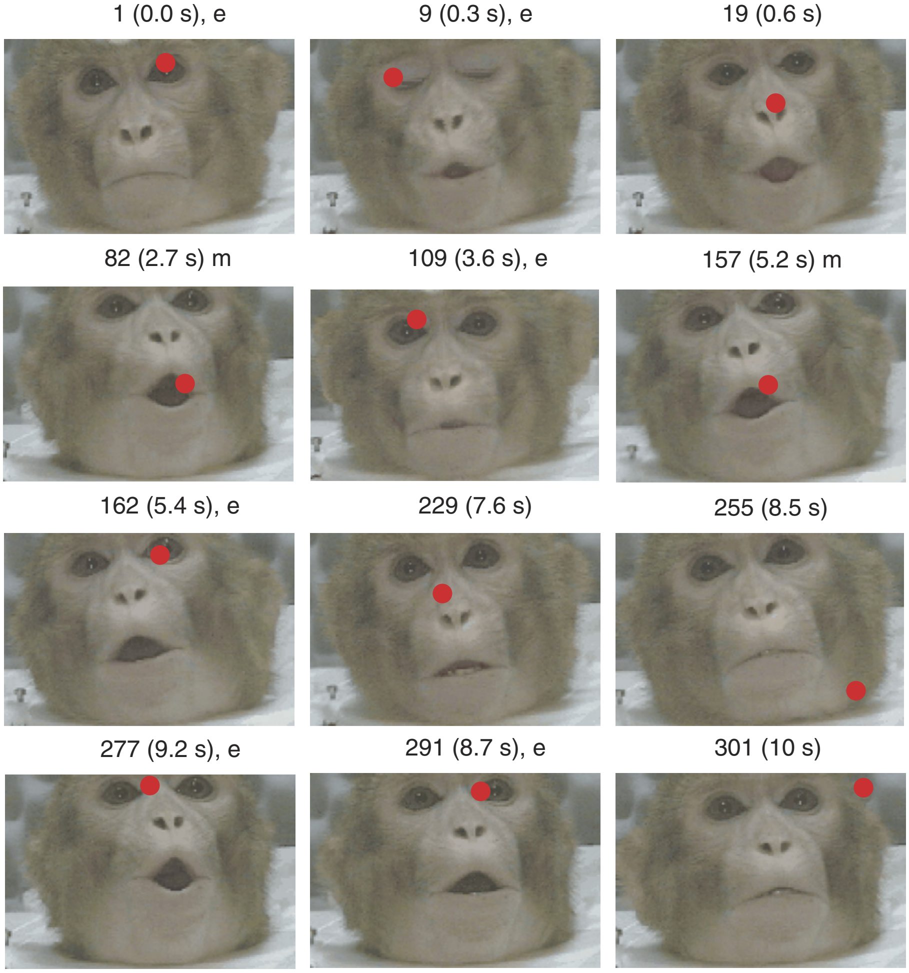

Fig. 2. Representative example of the fixation patterns of one subject monkey viewing Movie #1 under the 'equal noise' condition. Each fixation is represented by a dot. Only frames on which fixations fell upon the face are shown here. Almost all fixations fell in the eye region, but occasional fixations occurred on the mouth when it opened. Frame numbers are given above each frame along with the equivalent duration in seconds, and the letters (e, eye; m, mouth) indicate which face region the fixation fell according to our criteria.

Although monkeys looked at the mouth considerably less than the eyes during the video presentations, their fixations on the mouth were tightly correlated with mouth movements. Fig. 5 shows the correlation between the onset of fixations on the mouth and the onset of the vocalizations ($r = 0.997$, $p < 0.0001$). It is possible that the distribution of mouth fixations varied according to call type and/or noise levels. To address this, we measured the number of fixations on the mouth per call type relative to the total number of fixations on the mouth. An ANOVA revealed that there were no differences

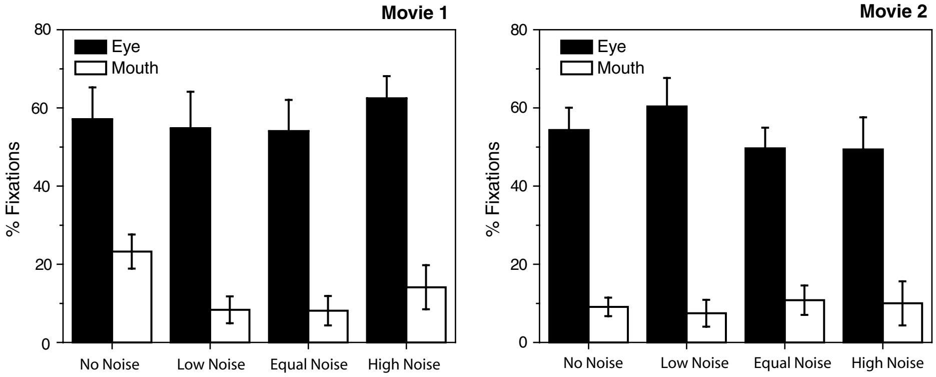

Fig. 3. The average fixation on the eye region versus the mouth region across all four subjects. Background noise had little or no influence on the proportion of fixations falling onto the mouth or the eye region. This outcome was not influenced by the movie type. Error bars represent SEM.

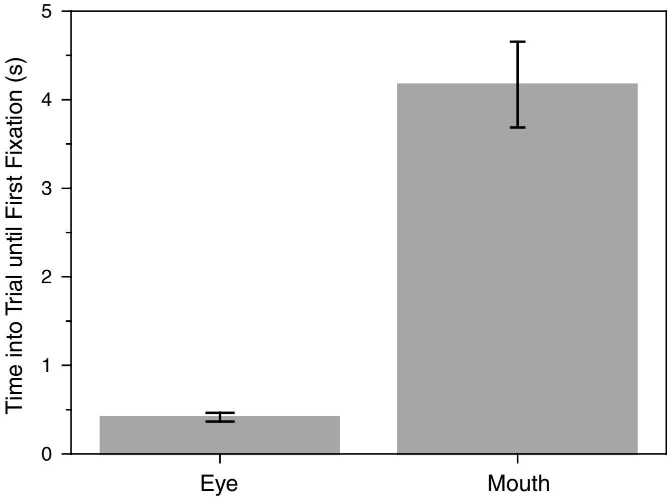

Fig. 4. Time into trial for the first fixation falling onto eye or mouth region. Whereas the eyes were fixated within the first 500 ms of the trial, fixations on the mouth happened only after several seconds into the movie. Error bars represent SEM.

in the frequency of mouth fixations between call types in either movie sequence (Movie 1: $F(1,3) = 2.47$, $p = 0.21$; Movie 2: $F(2,6) = 1.85$, $p = 0.24$) or across the different noise conditions (Movie 1: $F(3,9) = 0.36, p = 0.78$; Movie 2: $F(3,9) = 1.0, p = 0.44$). Another possibility is that the duration of fixations varied according to call type or noise levels. However, again, there were no significant differences across different noise levels (for both movies together: $F(3,18) = 0.15, p = 0.93$).

Overall, this pattern of results leaves open the possibility that the auditory component of the vocalizations has *no* influence on the eye movements of rhesus monkey

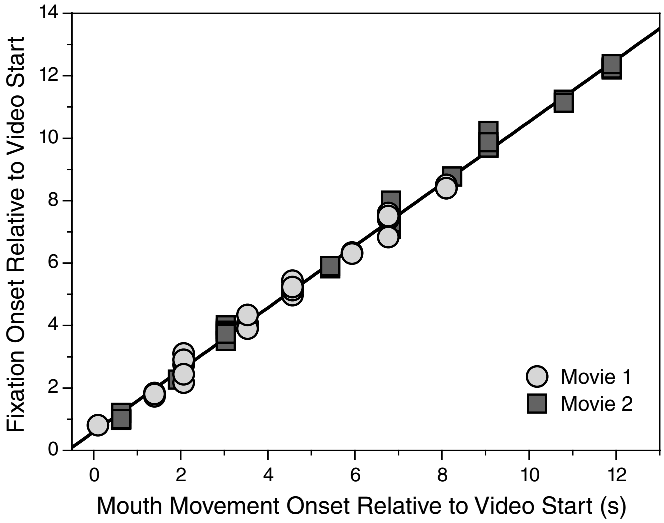

Fig. 5. Fixations in the mouth region occur almost invariably when there is a detectable movement from that region. This figure shows the tight correlation between mouth fixation onsets and the onset of mouth movements in the video.

observers. Furthermore, it is possible that the abrupt frame transitions between the vocalization clips spliced together in the movies obscured eye movement patterns that would occur in a more natural video sequence. We therefore ran a second experiment with Movie 3 on three of the four monkey subjects to examine these possibilities. The video sequence in this experiment was 30 s in duration and of a single novel individual producing three different call types in a single recording session: a grunt, three coos, and a scream (Fig. 6A). Minimal editing was done to this video so there

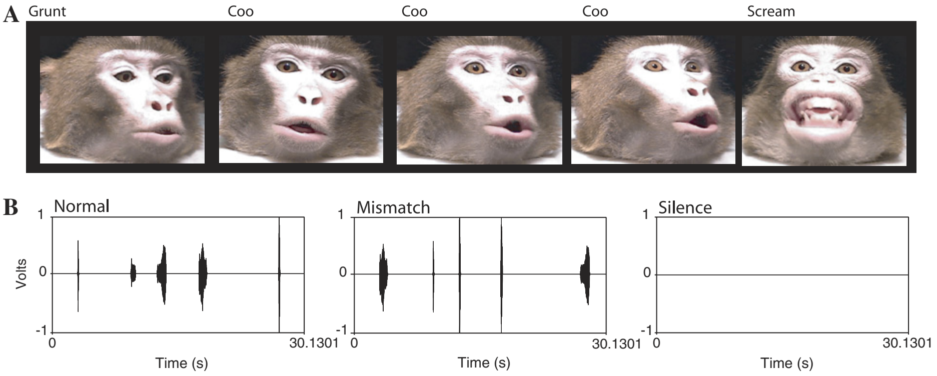

Fig. 6. The stimuli used for the Normal versus Mismatch versus Silent conditions. (A) Frames from the midpoint of each call type. (B) Time-waveforms of the audio tracks for the Normal (top), Mismatch (middle), and Silent (bottom) conditions.

are long periods of neutral expressions or lipsmack expressions between the vocalizations and only two abrupt frame transitions which occurred during neutral postures. Three conditions were tested with this video sequence: (1) Normal (audio track as originally recorded), (2) Mismatch (call types shuffled but retain the appropriate onset times), and (3) Silence (Fig. 6B). If the auditory component of these vocalizations influences eye movement patterns, then the proportion of fixations should differ between the normal and mismatch/silence conditions.

Contrary to our expectations, there was no influence of audition on the eye movement patterns of rhesus monkeys viewing vocalizing conspecifics. An ANOVA revealed that there were no significant differences between the Normal, Mismatch, and Silence conditions ($F(2,12) = 0.208$, $p = 0.815$), but there was a significant difference between the proportion of fixations falling on the eyes versus the mouth ($F(1,12) = 24.29$, $p < 0.0001$) (Fig. 7). As in the prior experiment with background noise, fixations on the mouth were correlated with the onset of mouth movements ($r = 0.96$, $p < 0.0001$). However, now the influence of the auditory component can be ruled out as this pattern of mouth fixations occurred even when there was no audio track (silent condition only: $r = 0.93$, $p < 00001$) (Fig. 8).

4. Discussion

We allowed rhesus monkeys to freely view video sequences of conspecific individuals producing vocalizations. Under all listening conditions, our monkey subjects spent most of their time inspecting the eye region relative to the mouth. When they did fixate on the mouth, it was highly correlated with the onset of mouth movements. Finally, there was no relationship between the number or duration of fixations with respect to call type. We conclude, therefore, that the

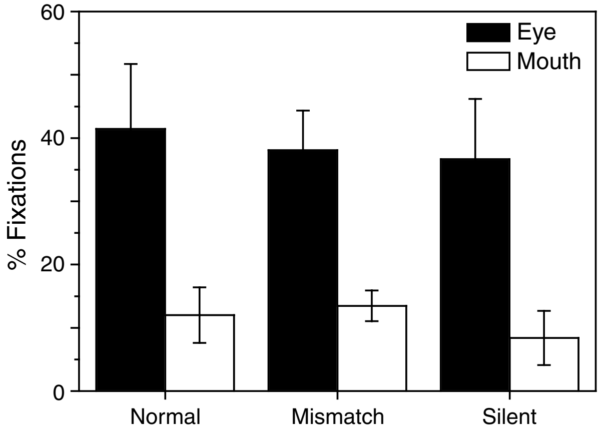

Fig. 7. The average fixation on the eye region versus the mouth region across three subjects. The audio track had no influence on the proportion of fixations falling onto the mouth or the eye region. Error bars represent SEM.

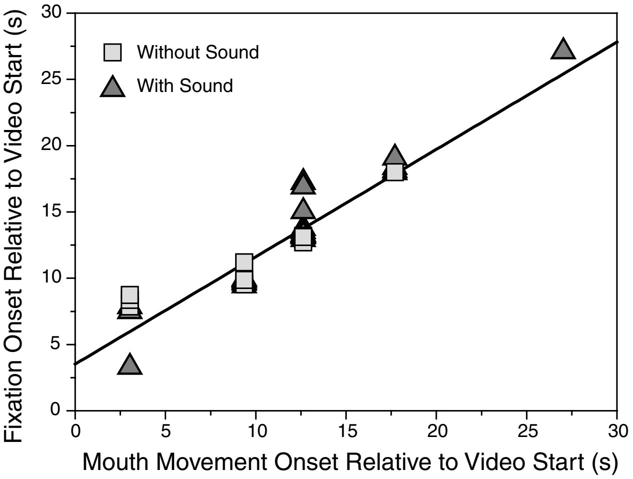

Fig. 8. As in the background noise conditions, fixations in the mouth region occur almost invariably when there is a detectable movement from that region. This figure shows that the tight correlation between mouth fixation onsets and the onset of mouth movements in the video is independent of the audio track.

auditory component has no influence on eye movement patterns of monkeys viewing vocalizing conspecifics.

Nonetheless, our findings have striking parallels with what we know about human eye movement patterns during speech-reading. In both species, the greater number of fixations fall in the eye region than in the mouth region when subjects are required simply to view vocalizing conspecifics (Klin et al., 2005), to attend to emotion-related cues or to make social judgments (Buchan et al., 2004, 2005). Even during visual speech alone (no auditory component), when subjects are asked to attend to prosodic cues, they will look at the eyes more than the mouth (Lansing & McConkie, 1999). Furthermore, like human observers (Lansing & McConkie, 2003), monkeys look at the eyes *before* they look at the mouth and their fixations on the mouth are tightly correlated with mouth movement. For instance, Lansing and McConkie (2003) reported that, regardless of whether it was visual or audiovisual speech, subjects asked to identify words increased their fixations onto the mouth region with the onset of facial motion.

Our monkey data diverge from the human condition when considering the influence of noise. When humans are required to identify speech-specific signals, then they will increase their fixations on the mouth relative to the eyes. For example, when asked to identify words or phonetic information embedded in audiovisual speech in the presence of background noise (Vatikiotis-Bateson et al., 1998) or under ideal listening conditions (Buchan et al., 2004, 2005), subjects will fixate on the mouth region more than the eye region. In contrast, monkeys appear to always prefer the eye region regardless of the noise conditions. Indeed, under conditions where the background noise is extremely high or where the auditory component mismatches the visual component, monkeys continue to focus on the eyes more than the mouth. There two possible explanations for this. One possibility is that the natural viewing mode of monkeys is to extract social information and emotional content from vocal communication signals and thus their eye movement patterns are similar to the eye movement patterns of humans during natural viewing contexts (Klin et al., 2005) and during tasks related to emotional or social judgments (Buchan et al., 2004, 2005; Lansing & McConkie, 1999). The other possibility is that the monkeys have adopted a species-*atypical* behavior and/or they would have radically different eye movement patterns if required to identify auditory content in a task of some sort.

We did not train the monkeys to report what they heard. It is possible that when forced to use auditory information in order to receive a reward that they would then modify their eye movements to maximize their ability to discriminate target calls from the background noise. Such modifications may include looking at the mouth more frequently. While monkeys can be trained to perform a variety of complex tasks, such training hampers interpretations regarding the *natural* capacities of monkeys and their neural mechanisms. As we are interested in the behavior patterns of monkeys that would allow us to make explicit claims about their species-typical capacities, training the animals to perform an identification task would have defeated the purpose of the comparison with humans. As suggested above, another possible explanation for our data is that the subjects were using unusual eye movement patterns given the pseudo-natural context of the experiments. We think it is unlikely that rhesus monkeys are using a different, novel strategy (that is, totally ignoring the auditory component) than they would in their natural contexts for the following reason. Using a preferential looking paradigm, we have previously shown that rhesus monkeys can spontaneously (i.e., without training or reward) attend to the correct facial expression when they hear a corresponding conspecific vocalization (Ghazanfar & Logothetis, 2003) and that they can match the number of voices they concurrently hear with the number of faces they see – an ability that requires them to segregate up to three conspecific voices in the auditory domain and then preferentially attend to the correct visual display (Jordan, Brannon, Logothetis, & Ghazanfar, 2005). These experiments were also conducted in an artificial context (darkened room with video monitors) while the monkey subject sat in a primate chair. Thus, monkeys are able to match faces and voices spontaneously in an experimental context; there is no reason to think that they would do so in a manner that is wholly different than the one they use in real social contexts.

If the eye movement strategies of monkeys viewing vocalizing faces are primarily to glean social information, then why fixate on the eyes? Many previous experiments have shown that monkeys prefer to look at the eyes when viewing neutral or expressive faces (Guo, Robertson, Mahmoodi, Tadmor, & Young, 2003; Keating & Keating, 1982; Nahm, Perret, Amaral, & Albright, 1997) and the attention directed at the eyes often seems to be used to assess the intention of a conspecific or other competitor (Ghazanfar & Santos, 2004). Indeed, a number of primate species, including rhesus macaques, will spontaneously orient to where other individuals are looking (Deaner & Platt, 2003; Emery, Lorincz, Perrett, Oram, & Baker, 1997; Tomasello, Call, & Hare, 1998) and use this information to adapt their behavior (Flombaum & Santos, 2005). Thus, monkeys may focus on the eyes when observing a conspecific's vocalization to glean information about his/her intentions from the most accurate source.

The eye region may also give clues about what is occurring in the mouth region and therefore eliminate the need to always look directly at the mouth to know its posture. As proposed by Vatikiotis-Bateson et al. (1998) for humans, it is possible that perceivers acquire vocalization-related information that is distributed broadly on the vocalizer's face. Facial motion during speech is a direct consequence of the vocal tract movements necessary to shape the acoustics of speech – indeed, a large portion of the variance observed in vocal tract motion can be estimated from facial motion (Yehia, Kuratate, & Vatikiotis-Bateson, 2002). Humans, therefore, can identify vocal sounds when the mouth is masked or without directly looking at the mouth presumably by using such facial motion cues (Preminger, Lin, Payen, & Levitt, 1998). Head movement can also be an informative cue, one that is linked to the fundamental frequency (*F*0) and voice amplitude of the speech signal (Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004; Yehia et al., 2002). When head movements are eliminated or distorted in speech displays, speech perception is degraded (Munhall et al., 2004). Another possibility is that saccades to the mouth are a reflexive response to motion detection in the visual periphery (Vatikiotis-Bateson et al., 1998). Thus, for many of the same reasons applied to human perceivers, rhesus monkeys may simply not need to always look at the mouth to know which facial posture accompanies a vocalization.

As in humans, different rhesus monkey vocalizations are produced with unique facial expressions and the motion of articulators influences the acoustics of the signal (Hauser et al., 1993; Hauser & Ybarra, 1994). Such articulatory postures could potentially influence facial motion beyond the mouth region. For example, grimaces produced during scream vocalizations cause the skin folds around the eyes to increase in number (Hauser, 1993). In addition to these production-related facial movements, some vocalizations are associated with visual cues that are not directly related to the articulatory movement. Threat vocalizations, for instance, are produced with intense staring, eyebrows raised, and ears often pulled back (Partan, 2002). Head position and motion (e.g., chin up versus chin down versus neutral position) also vary according to vocal expression type (Partan, 2002). Thus, it is likely that many of the facial motion cues that humans use for speechreading are present in rhesus monkeys as well.

In conclusion, our data suggest that, in large part, monkeys and humans share homologous sensorimotor strategies when processing bimodal vocal communication signals. In both species, the eye region of a conspecific's face is more important than the mouth region. In humans, this is particular true in contexts where extracting social or emotional information is a priority. As monkeys do not have anything akin to 'words', there is no need to extract speech-like information, and thus their eye movements are in accord with the notion that they are extracting social/emotional information. Furthermore, the overall eye movement patterns of monkeys are consistent with the two-force model put forth by Lansing and McConkie (2003) for humans: one force that draws attention to the eyes for social reasons and another that draws attention to the mouth when there is vocalization-associated movement.

Acknowledgments

We thank Joost Maier and Kari Hoffman for their comments on a previous version of this manuscript and two anonymous reviewers for their thoughtful suggestions. This work was supported by the Max Planck Society.

References

Andrew, R. J. (1962). The origin and evolution of the calls and facial expressions of the primates. *Behaviour, 20*, 1–109.

Brown, C. H. (2003). Ecological and physiological constraints for primate vocal communication. In A. A. Ghazanfar (Ed.), *Primate audition: Ethology and neurobiology* (pp. 127–150). Boca Raton, FL: CRC Press.

Buchan, J. N., Pare, M., & Munhall, K. G. (2004). The influence of task on gaze during audiovisual speech perception. *Journal of the Acoustical Society of America, 115*, 2607.

Buchan, J. N., Pare, M., & Munhall, K. G. (2005). Gaze behavior during the processing of dynamic faces. *Society for Neuroscience Abstracts.*

Cotton, J. C. (1935). Normal "visual hearing". *Science, 82*, 592–593.

Deaner, R. O., & Platt, M. L. (2003). Reflexive social attention in monkeys and humans. *Current Biology, 13*(18), 1609–1613.

Emery, N. J., Lorincz, E. N., Perrett, D. I., Oram, M. W., & Baker, C. I. (1997). Gaze following and joint attention in rhesus monkeys (*Macaca mulatta*). *Journal of Comparative Psychology, 111*(3), 286–293.

Flombaum, J. I., & Santos, L. R. (2005). Rhesus monkeys attribute perceptions to others. *Current Biology, 15*(5), 447–452.

Ghazanfar, A. A., & Logothetis, N. K. (2003). Facial expressions linked to monkey calls. *Nature, 423*(6943), 937–938.

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience, 25*, 5004–5012.

Ghazanfar, A. A., Neuhoff, J. G., & Logothetis, N. K. (2002). Auditory looming perception in rhesus monkeys. *Proceedings of the National Academy of Sciences of the United States of America, 99*(24), 15755–15757.

Ghazanfar, A. A., & Santos, L. R. (2004). Primate brains in the wild: The sensory bases for social interactions. *Nature Reviews Neuroscience, 5*(8), 603–616.

Ghazanfar, A. A., Smith-Rohrberg, D., Pollen, A. A., & Hauser, M. D. (2002). Temporal cues in the antiphonal long-calling behaviour of cottontop tamarins. *Animal Behaviour, 64*, 427–438.

Guo, K., Robertson, R. G., Mahmoodi, S., Tadmor, Y., & Young, M. P. (2003). How do monkeys view faces? – A study of eye movements. *Experimental Brain Research, 150*(3), 363–374.

Hauser, M. D. (1993). Right-hemisphere dominance for the production of facial expression in monkeys. *Science, 261*(5120), 475–477.

Hauser, M. D., Evans, C. S., & Marler, P. (1993). The role of articulation in the production of rhesus-monkey, *Macaca mulatta*, vocalizations. *Animal Behaviour, 45*(3), 423–433.

Hauser, M. D., & Marler, P. (1993). Food-associated calls in rhesus macaques (*Macaca mulatta*).1. Socioecological factors. *Behavioral Ecology, 4*(3), 194–205.

Hauser, M. D., & Ybarra, M. S. (1994). The role of lip configuration in monkey vocalizations – Experiments using xylocaine as a nerve blocks. *Brain and Language, 46*(2), 232–244.

Hinde, R. A., & Rowell, T. E. (1962). Communication by posture and facial expressions in the rhesus monkey (*Macaca mulatta*). *Proceedings of the Zoological Society London, 138*, 1–21.

Jordan, K. E., Brannon, E. M., Logothetis, N. K., & Ghazanfar, A. A. (2005). Monkeys match the number of voices they with the number of faces they see. *Current Biology, 15*, 1034–1038.

Keating, C. F., & Keating, E. G. (1982). Visual scan patterns of rhesus monkeys viewing faces. *Perception, 11*, 211–219.

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2005). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry, 59*, 809–816.

Kuhl, P. K., Williams, K. A., & Meltzoff, A. N. (1991). Cross-modal speech perception in adults and infants using nonspeech auditory stimuli. *Journal of Experimental Psychology: Human perception and performance, 17*, 829–840.

Lansing, C. R., & McConkie, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech Language and Hearing Research, 42*(3), 526–539.

Lansing, I. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception & Psychophysics, 65*(4), 536–552.

Maier, J. X., Neuhoff, J. G., Logothetis, N. K., & Ghazanfar, A. A. (2004). Multisensory integration of looming signals by rhesus monkeys. *Neuron, 43*(2), 177–181.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility – Head movement improves auditory speech perception. *Psychological Science, 15*(2), 133–137.

Nahm, F. K. D., Perret, A., Amaral, D. G., & Albright, T. D. (1997). How do monkeys look at faces?. *Journal of Cognitive Neuroscience 9*, 611–623.

Partan, S. R. (2002). Single and multichannel signal composition: Facial expressions and vocalizations of rhesus macaques (*Macaca mulatta*). *Behaviour, 139*, 993–1027.

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science, 6*(2), 191–196.

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

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lipreading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97–113). Hillsdale, NJ: Erlbaum.

Rowe, C. (1999). Receiver psychology and the evolution of multicomponent signals. *Animal Behaviour, 58*, 921–931.

Rowell, T. E., & Hinde, R. A. (1962). Vocal communication by the rhesus monkey (*Macaca mulatta*). *Proceedings of the Zoological Society London, 138*, 279–294.

Rudmann, D. S., McCarley, J. S., & Kramer, A. F. (2003). Bimodal displays improve speech comprehension in environments with multiple speakers. *Human Factors, 45*, 329–336.

Shannon, R. V., Zeng, G. F., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*, 303–304.

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

Tomasello, M., Call, J., & Hare, B. (1998). Five primate species follow the visual gaze of conspecifics. *Animal Behaviour, 55*, 1063–1069.

Van Hooff, J. A. R. A. M. (1962). Facial expressions of higher primates. *Symposium of the Zoological Society, London, 8*, 97–125.

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics, 60*(6), 926–940.

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics, 30*(3), 555–568.