Integration of Visual and Auditory Information by Superior Temporal Sulcus Neurons Responsive to the Sight of Actions

Nick E. Barraclough[1,*], Dengke Xiao[1,*], Chris I. Baker[2], Mike W. Oram[1], and David I. Perrett[1]

Abstract

■ Processing of complex visual stimuli comprising facial movements, hand actions, and body movements is known to occur in the superior temporal sulcus (STS) of humans and nonhuman primates. The STS is also thought to play a role in the integration of multimodal sensory input. We investigated whether STS neurons coding the sight of actions also integrated the sound of those actions. For 23% of neurons responsive to the sight of an action, the sound of that action significantly modulated the visual response. The sound of the action increased or decreased the visually evoked response for an equal number of neurons. In the neurons whose visual response was increased by the addition of sound (but not those neurons whose responses were decreased), the audiovisual integration was dependent upon the sound of the action matching the sight of the action. These results suggest that neurons in the STS form multisensory representations of observed actions. ■

INTRODUCTION

Many actions make a noise. The movement of faces is often associated with vocalizations and sounds of food being ingested, walking with the sound of footsteps, and hand actions with different sounds depending upon the interaction with objects (e.g., tearing, hitting, or manipulating). Although the addition of auditory information normally associated with the sight of an action can help us interpret and understand that action better (Dodd, 1977), the addition of auditory information incongruent with the visual information can disrupt our ability to form a coherent percept of social signals (McGurk & MacDonald, 1976). Thus, there are psychologically important interactions due to the matching of auditory and visual processing.

Matching of visual and auditory signals is apparent in motor circuits. "Mirror neurons" in the rhesus macaque premotor cortex appear to match motor circuits for generating facial communicative gestures and hand actions with incoming visual signals about the same actions (Ferrari, Gallese, Rizzolatti, & Fogassi, 2003; Gallese, Fadiga, Fogassi, & Rizzolatti, 1996; di Pelligrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992). Matching of motor and sensory representations in the premotor cortex extends to auditory signals as well as visual input (Kohler et al., 2002). Understanding actions could be based upon a matching of sensory input to the motor representation of the same action within premotor cortex (Bekkering & Wohlschlaeger, 2002; Prinz, 2002; Viviani, 2002; Iacoboni et al., 1999). Alternatively, a more feed-forward process could occur whereby a polysensory representation of the action is formed before matching to motor circuits. Integration of visual and auditory signals within polysensory neurons could form a basis for understanding the action without reference to motor circuits involved in the generation of that action.

The human superior temporal sulcus (STS) has been implicated as a region where multisensory integration may occur, not just with signals received from overlapping regions of sensory space but also with reference to the content of the information that the sensory signals might carry (Calvert, 2001). Calvert (2000) examined human brain activity to a moving face talking (the visual input) and speech (the auditory input). The "phoneme" structure of the sound of speech was either congruent or incongruent with the "viseme" structure of the sight of the speaking face. With congruent visual and auditory information, the blood oxygen level-dependent (BOLD) response from the ventral bank of the left STS was greater than a response predicted by the addition of the responses to the visual and auditory stimuli presented separately (supra-additive). A posterior region of the human STS also shows a supra-additive BOLD response to visual and auditory signals

[1]University of St Andrews, Scotland, [2]Massachusetts Institute of Technology
*These authors contributed equally to the work.

© 2005 Massachusetts Institute of Technology

when presented with combined audiovisual stimuli of objects (e.g., telephone) making noises (Beauchamp, Lee, Argall, & Martin, 2004).

Human STS may be involved in audiovisual integration of social and complex object cues but not in a general audiovisual integration per se. Bushara et al. (2003) used a visual presentation of two moving bars converging, passing over each other, and continuing their trajectory. A collision sound was played as the two bars touched. The collision sound increased the likelihood of perceiving the bars rebounding rather than seeing them as passing over each other. BOLD activity during the perception of the collision was increased over premotor, superior colliculus, insula, and superior parietal cortex but not in the temporal lobe.

The results from imaging studies indicate cortical loci where integration occurs. How integration is performed within these areas is largely unexplained by imaging. The majority of physiological studies of audiovisual integration using simple visual and auditory stimuli have concentrated upon the cat and primate superior colliculus (Bell, Corneil, Munoz, & Meredith, 2003; Wallace, Wilkinson, & Stein, 1996; Jay & Sparks, 1984; Meredith & Stein, 1983, 1986a, 1986b, 1996). These studies have illustrated several of the underlying response properties of neurons showing audiovisual integration.

Neurons showing audiovisual integration can do more than respond to the different unimodal stimuli, the response to the multimodal stimulus is often greater than the sum of the responses to the unimodal stimuli presented alone, a supra-additive response. Neurons in the superior colliculus that show multisensory integration have overlapping visual and auditory receptive fields and can show supra-additive responses when the audiovisual stimuli are presented in that same region of space (Meredith & Stein, 1986a, 1996). If either the auditory or visual stimulus is presented outside the neuron's auditory or visual receptive field, there is either no supra-additive response or there is an inhibition of the neuronal response (Kadunce, Vaughan, Wallace, Benedek, & Stein, 1997; Meredith & Stein, 1986a, 1996). Furthermore, the visual and auditory stimuli must be in close temporal contiguity for the neuronal responses to show supra-additivity (see Stein & Meredith, 1993, for a review).

Single-cell physiology in macaque STS has shown that cells can code both visual and auditory information in the upper bank and fundus of the STS (Hikosaka, Iwai, Saito, & Tanaka, 1988; Bruce, Desimone, & Gross, 1981) and also in the lower bank of the STS (Benevento, Fallon, Davis, & Rezak, 1977). Benevento et al. (1977) estimated that the proportion of neurons in both banks of the STS that have both auditory and visual responses was about 36%; Bruce et al. (1981) reported about 38% in the upper bank and fundus. Hikosaka et al. (1988) reported that 12% of neurons in a more caudal region of the STS responded to both auditory and visual stimuli. Little is known, however, about what auditory and visual information might be integrated or the underlying integrative mechanisms. These early studies showed that STS neurons that respond to the visual presentation of hands, faces, and moving objects could also respond to beeps, clicks, white noise, or voices (Bruce et al., 1981; Benevento et al., 1977). Although audiovisual interactions were not systematically studied, there was evidence for sound attenuating visual responses and some neurons responding only to combined audiovisual stimuli.

Human functional magnetic resonance imaging (fMRI) data (Calvert, 2000) suggests that STS cells might combine unimodal inputs based upon similarity of higher order statistics of the auditory and visual stimuli rather than similarity of the spatial and temporal characteristics. Consistent with results from human imaging studies of visual processing of actions (Puce & Perrett, 2003; Downing, Jiang, Shuman, & Kanwisher, 2001; Grossman et al., 2000; Puce, Allison, Bentin, Gore, & McCarthy, 1998), monkey STS contains neurons that respond to the appearance of complex visual stimuli. Many cells in the upper bank, lower bank, and the fundus of the STS respond to bodies walking (Oram & Perrett, 1994, 1996; Perrett et al., 1985); cells in the lower bank and fundus of the STS have also been shown to respond to hands grasping and manipulating objects (Perrett, Mistlin, Harries, & Chitty, 1990), and other cells in the upper bank, lower bank, and fundus of the STS respond to pictures of different facial expressions (Sugase, Yamane, Ueno, & Kawano, 1999; Hasselmo, Rolls, & Baylis, 1989; Perrett et al., 1984).

Visual areas TE and TEO (Von Bonin & Bailey, 1947) and medial superior temporal (MST, Desimone & Ungerleider, 1986; Ungerleider & Desimone, 1986) project to the upper bank, lower bank and fundus of rostral STS (Saleem, Suzuki, Tanaka, & Hashikawa, 2000; Seltzer & Pandya, 1978, 1984). In addition, the upper bank and fundus of the STS receive input from auditory regions including the superior temporal gyrus (Seltzer & Pandya, 1978), a region containing neurons shown to respond selectively to different monkey calls (Rauschecker, Tian, & Hauser, 1995). The lower bank of the STS is interconnected with upper bank via the fundus (Seltzer & Pandya, 1989). In short, both visual and auditory information are available to the dorsal and ventral banks and fundus of rostral STS either directly or indirectly from neighboring areas.

In summary, the available data suggest that the STS could be involved in the binding of visual and auditory information about actions, including communicative signals, to provide a multisensory neural representation of actions. To test whether STS neurons responsive to the sight of actions integrate auditory information with visual information, we first searched for visually responsive STS cells that were particularly sensitive to the sight of actions. For each action, we then tested the cell

response sensitivity to a sound that would normally be associated with the action and the effect of that sound when paired with the visual stimulus. To determine whether the type of auditory input was important for sensory integration, we measured the responses to the action when incongruent sounds were included with the visual stimulus. We hypothesized that sound of action would affect STS cell responses to the sight of actions and that audiovisual congruence would matter.

RESULTS

We recorded from 545 single cells in the temporal lobe (upper and lower banks of the STS and IT) from 2 monkeys. One hundred seventy-four cells responded to auditory stimuli, consistent with previous studies showing auditory responses in STS cells (Hikosaka et al., 1988; Bruce et al., 1981; Benevento et al., 1977). Clinical testing for audiovisual interactions was performed in 20 auditory responsive cells. For 10 of the 20 auditory responsive cells (50%), the multisensory response was significantly different from the individual sensory responses, tested with ANOVA. Figure 1 summarizes two cell responses to the action of foot tapping. The cell in Figure 1A shows a strong response to the sound of the action, this response is significantly attenuated when the action was performed in sight. The cell in Figure 1B shows a significant augmentation of the response to a unimodal stimulus when the other unimodal stimulus is presented concurrently.

We focused on how the addition of auditory signals affected the response to visually presented actions. One hundred forty-seven neurons were responsive to different actions, facial movements, hand actions, or walking movements. Ninety-five were tested for the effect of auditory signals on the visual response.

Effect of Sound on Visual Responses

Auditory signals had a significant effect on the visual response in 22 (23%) of the 95 cells with visual responses. The visual response was significantly augmented in 8 of 95 cells and significantly attenuated in 8 of 95 cells. For an additional 6 of 95 cells, the addition of an auditory signal produced both a significant augmentation and a significant attenuation, depending upon the action tested. Here we treat the two sets of augmented and attenuated responses to the two different actions separately in these cells.

Figure 2 shows an example of augmentation of the visual response of a single cell by auditory signals. The mean response to the visual stimulus, in this case, human hands tearing a piece of paper, is 73 spikes/sec, increasing to 198 spikes/sec when the visual and auditory stimuli are combined. The response to the auditory stimulus alone is 10 spikes/sec. The percentage increase in the visual response with the addition of the auditory signal (taking into account the background firing rate) is 231%, the index of the linearity of integration (I_{audvis}, see Methods) is 3.0, indicating that the augmentation of the visual response by the addition of auditory signals is supra-additive.

An example attenuation of a visual response by auditory signals is illustrated in Figure 3. For this cell, the mean response to the visual stimulus, a human face chewing, is 22 spikes/sec. The response decreases to 13 spikes/sec when the sight and sound of the chewing

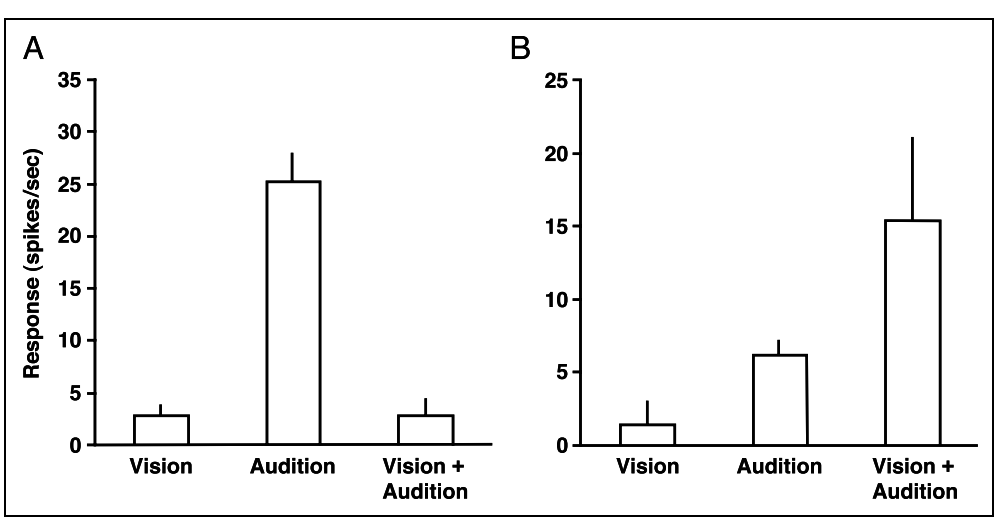

Figure 1. Responses of two STS cells showing audiovisual interaction. The stimuli consisted of the action of foot-tapping (not illustrated due to the clinical nature of the testing). Mean responses (± *SEM*) of two STS cells to visual, auditory, and combined visual and auditory stimuli are plotted. The response of cell A to an auditory stimulus (mean response ± *SEM*) is attenuated by the addition of a visual signal (mean response ± *SEM*) [ANOVA: $F(2,16) = 24.308$, $p < .0001$, PLSD post hoc test, $p < .05$]. The response of cell B to unimodal visual or auditory stimuli is augmented by the addition of the other unimodal input [ANOVA: $F(2,27) = 43.43$, $p < .0001$, PLSD post hoc test, $p < .05$, each comparison].

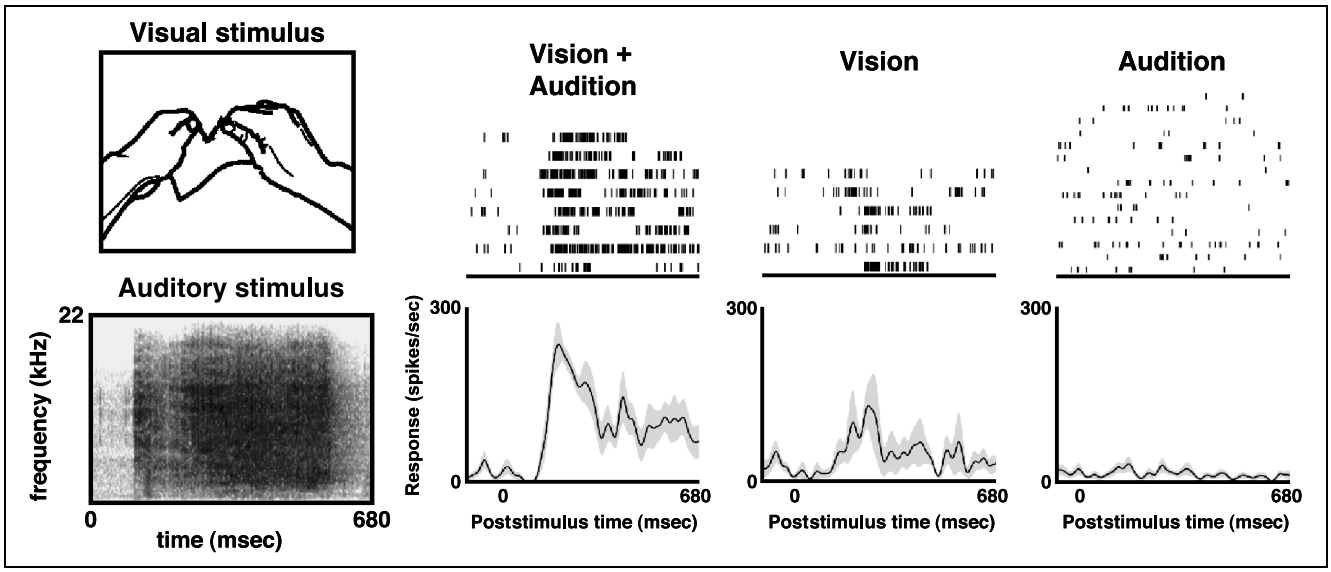

Figure 2. Augmentation of visual responses by an auditory signal. Top left: Illustration of the middle frame from a movie of hands tearing paper. Bottom left: Spectrogram of the auditory stimulus, *x*-axis = time (msec) and *y*-axis = frequency, (kHz), and amplitude represented by the darkness of the gray scale (white 35 dB, black 100 dB). Right: The responses of a single cell to a movie of hands tearing paper with both visual and auditory stimuli (trials = 8), with the visual stimulus alone (trials = 6), and with the auditory stimulus alone (trials = 14). The upper section of each plot shows individual trial responses as rastograms, the lower section the SDFs calculated from all trials (gray = *SEM*). The response to the combined visual and auditory stimuli was significantly larger than the response to the visual stimulus ($p < .05$).

are combined. The response to the sound of chewing alone is 0 spikes/sec. The decrease in the visual response with the addition of auditory signals is 43% ($I_{\text{audvis}} = -0.43$), indicating that the attenuation of the visual response by auditory signals is sub-additive.

Population Responses of Cells Showing a Modulatory Effect of Auditory Signals

Cells where the visual response is unaffected by auditory signals have a mean firing rate for the combined vision

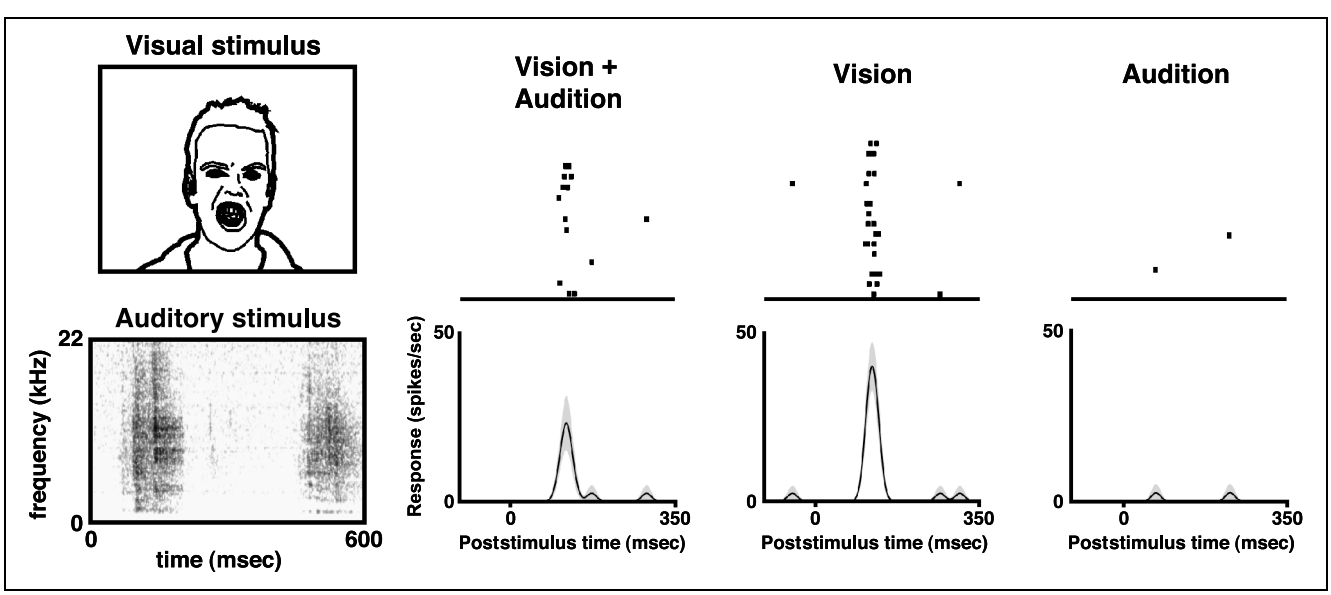

Figure 3. Attenuation of visual responses by an auditory signal. Top left: Illustration of the middle frame from a movie of a human chewing. Bottom left: Spectrogram of the auditory stimulus, *x*-axis = time (msec) and *y*-axis = frequency (kHz), and amplitude represented by the darkness of the gray scale, white 50 dB, black 100 dB). Right: Responses of a single cell to a movie of a human face chewing with both visual and auditory stimuli (trials = 17), with the visual stimulus alone (trials = 18), and with the auditory stimulus alone (trials = 16). The upper section of each plot shows individual trial responses as rastograms, the lower section the SDFs calculated from all trials (gray = *SEM*). The response to the combined visual and auditory stimuli was significantly smaller than the response to the visual stimulus ($p < .05$).

Table 1. Mean Responses and Latency Estimates of Cells Whose Visual Responses Were Augmented, Attenuated, or Unaffected by Auditory Signals

Cell Responses Classification	*V*				*VA*				*A*			
	Response (spikes/sec)		*Latency (msec)*		*Response (spikes/sec)*		*Latency (msec)*		*Response (spikes/sec)*		*Latency (msec)*	
	Mean (SD)	*n*	*Mean (SD)*	*n*	*Mean (SD)*	*n*	*Mean (SD)*	*n*	*Mean (SD)*	*n*	*Mean (SD)*	*n*
Augmentation	28.8 (26.1)	14	156 (61.8)	14	55 (51.8)	14	134 (73)	14	16.2 (17.6)	14	114.3 (72.1)	8
Attenuation	54.2 (60)	14	112.2 (68.7)	14	36.8 (46.4)	14	129 (47.8)	12	12.4 (21.8)	14	75 (37)	5
No Effect	24.6 (24)	73	109 (76)	73	23.4 (23)	73	115 (77)	71	11.7 (12)	73	115 (77)	52

Mean responses were calculated from the responses of each cell to the best action without subtraction of the background firing rate. Mean latencies are calculated from cell response onset latencies where detectable.
Standard deviations are given in parentheses.

and audition (VA) and vision-only (V) conditions in the region of 24 spikes/sec (see Table 1). For cells where the visual response is augmented by an auditory signal ($p < .05$), the mean firing rate is increased from 29 to 55 spikes/sec. Cells where the visual response is attenuated by an auditory signal have a mean firing rate of 54 spikes/sec to a visual stimulus, which drops to a mean of 37 spikes/sec with the addition of the auditory signal. For cells showing auditory augmentation of the visual response, the percentage change in the visual response with the addition of auditory signals ranged from 35% to 544% with a mean of 188%. The mean I_{audvis} was 1.24 (range −0.25 to 4.79), significantly larger than 0, $t(13) = 3.604$, $p < .005$, indicating that the integration of visual and auditory signals was, on average, supra-additive. For cells showing auditory attenuation of the visual response, the percentage change in the visual response with the addition of auditory signals ranged from a decrease of 22% to 131% with a mean decrease of 55%, the mean I_{audvis} was −0.54 (range −0.23 to −1.5), significantly less that 0, $t(13) = 6.068$, $p < .001$, indicating that the integration of visual and auditory signals was, on average, sub-additive.

Figure 4 shows the average spike density functions (SDFs) for the VA, V, and audition-only (A) conditions of the neurons with a visual response augmentation by the addition of auditory signals. Responses to the VA, V, and A conditions are shifted so that the contributing cells' V condition response onset latencies are time aligned at 100 msec. The magnitude of the visual response with the inclusion of an auditory stimulus increases by 86% of the response to the visual stimulus alone. There is little response to the auditory stimulus presented alone, and $I_{\text{audvis}} = 0.52$, indicating that the response to the com-

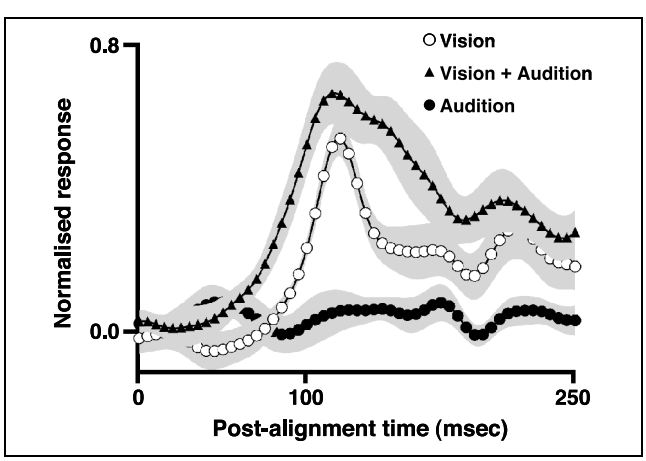

Figure 4. Responses of the average cell whose visual response is augmented by the addition of an auditory signal. Responses to combined visual and auditory stimuli (solid triangle), the visual stimulus alone (open circle), and the auditory stimulus alone (solid circle), averaged over 14 neurons in which auditory signals augment the visual response, are plotted as SDFs (gray = *SEM*).

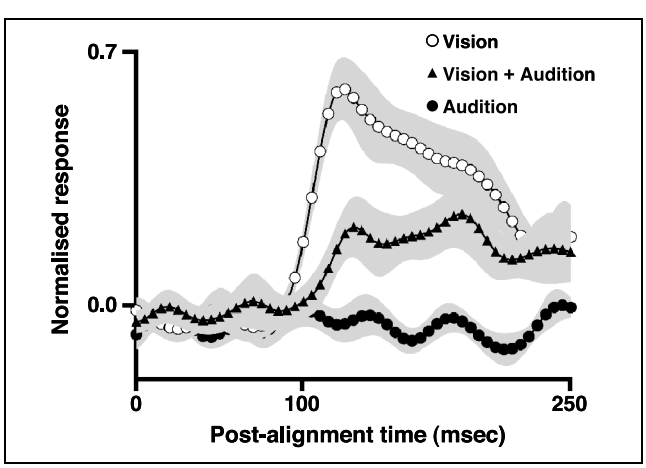

Figure 5. Responses of the average cell whose visual response is attenuated by the addition of an auditory signal. Responses to combined visual and auditory stimuli (solid triangle), the visual stimulus alone (open circle), and the auditory stimulus alone (solid circle), averaged over 14 neurons in which auditory signals attenuate the visual response, are plotted as SDFs (gray = *SEM*).

bined visual and auditory stimulus results from a nonlinear, supra-additive integration of the separate visual and auditory signals.

The responses of the average cell with visual response attenuation are plotted in a similar manner (Figure 5) showing the responses to the VA, V, and A conditions shifted so that the contributing cells' response onset latencies in the V condition are time aligned at 100 msec. The magnitude of the visual response with the inclusion of an auditory stimulus decreases by 46% of the response to the visual stimulus alone. There is little response to the auditory stimulus presented alone, and $I_{\text{audvis}} = -0.48$, indicating that the response to the combined visual and auditory stimulus results from a nonlinear, sub-additive integration of the separate visual and auditory signals.

Visual response latencies to the V and VA conditions are similar in cells with no effect of auditory signals on the visual response, $t(70) = 0.107, p > .9$ (see Table 1). For cells with a visual response augmentation by the addition of an auditory signal, visual response latencies are, on average, 156 msec. For these cells, the addition of auditory signals results in significantly earlier, $t(13) = 3.069$, $p < .01$, visual response latencies, on average, 134 msec. In the plot (Figure 4) of the average cell with a visual response augmentation, the earlier VA response latency is evident. As mentioned in the Methods section, the visual stimuli lag the auditory stimuli by 14–28 msec. It is, therefore, prudent to defer interpretation of the latency shift observed here until experiments are performed with a controlled range of time lags and leads. For cells showing visual response attenuation with the addition of an auditory signal, visual response latencies are 112 msec on average. In these cells, addition of auditory signals results in a later, on average, 129 msec, visual response latency, although this difference is not significant, $t(11) = 1.535, p = .153$. For the average cell with a visual response attenuation (Figure 5), the later VA response latency is also evident.

Effect of Type of Sound

Figure 6 shows responses of a single neuron that responds to the visual presentation of a human face lip-smacking (Figure 6A). The visual and congruent auditory stimulus combined (Figure 6B) produces a significant increase over the response to the visual stimulus alone (Figure 6A, $p < .05$). The responses to the incongruent combinations of the visual stimulus with different auditory vocalizations, a pant-threat (Figure 6C), and a coo (Figure 6D) are significantly smaller than the response to the congruent audiovisual combination ($p < .05$ each comparison) and not different from the visual stimulus presented alone ($p > .05$ each comparison). These results indicate that, for this neuron, the augmentation of the visual response by an auditory signal is dependent upon the nature of the auditory stimulus.

The integration of the visual signal and incongruent auditory signal was tested in 7 of 14 cells whose visual responses were augmented by congruent auditory signals. Figure 7 shows the average responses to the VA and the combined visual stimulus and incongruent auditory stimulus (VAi) conditions. For each of the seven cells, the responses to the VA stimuli are time aligned at 100 msec; each cell's corresponding VAi responses are shifted equivalent amounts. The average response to the combined visual stimulus and congruent auditory stimulus is significantly larger [ANOVA: $F(2,12) = 15.53$, $p < .0005$, Bonferroni-corrected protected least significant difference (PLSD), $p < .05$] than the average responses to the combined visual stimulus and incongruent auditory stimulus; responses were measured from 100 msec after time alignment. Although there was some augmentation of the visual response with incongruent auditory stimuli (24 vs. 26.1 spikes/sec), the data show response augmentation is, on average, significantly larger (180%) when the auditory stimuli "match" the presented visual stimuli (24 vs. 27.8 spikes/sec, $p < .05$).

The integration of the visual signal and incongruent auditory signal was tested in 8 of 14 cells whose visual responses were attenuated by the addition of congruent auditory signals. Calculated in the same way as above, the average response to the combined visual stimulus and congruent auditory stimulus is not significantly different ($p > .05$) than the average response to the combined visual stimulus and incongruent auditory stimulus.

In both Monkeys 1 and 2, cells showing responses to actions were found in the target area of the upper bank, lower bank, and fundus of rostral STS. As defined in previous studies (Saleem et al., 2000; Seltzer & Pandya, 1994; Distler, Boussaoud, Desimone, & Ungerleider, 1993; Hikosaka et al., 1988; Baylis, Rolls, & Leonard, 1987; Bruce, Desimone, & Gross, 1986; Bruce et al., 1981; Desimone & Gross, 1979), rostral STS is the region of cortex in the upper bank (TAa, TPO), lower bank (TEa, TEm), and fundus (PGa, IPa) of the STS that lies rostral to the fundus of the superior temporal sulcus (FST, Desimone & Ungerleider, 1986; Ungerleider & Desimone, 1986). The anterior–posterior extent of the recorded cells was from 7 to 10 mm anterior of the interaural plane, consistent with previous studies showing visual responses to actions in this region (Oram & Perrett, 1994, 1996; Perrett et al., 1989; Bruce et al., 1981). Overall, the majority of cells (30/32, 94%) showing an interaction of auditory and visual stimuli were located in the upper bank, lower bank, and fundus of rostral STS; the two additional cells were recorded in TE. It is important to note that cells showing an interaction of auditory and visual stimuli were located in all the regions of the rostral STS intermingled with cells that showed no interaction of auditory and visual stimuli in both

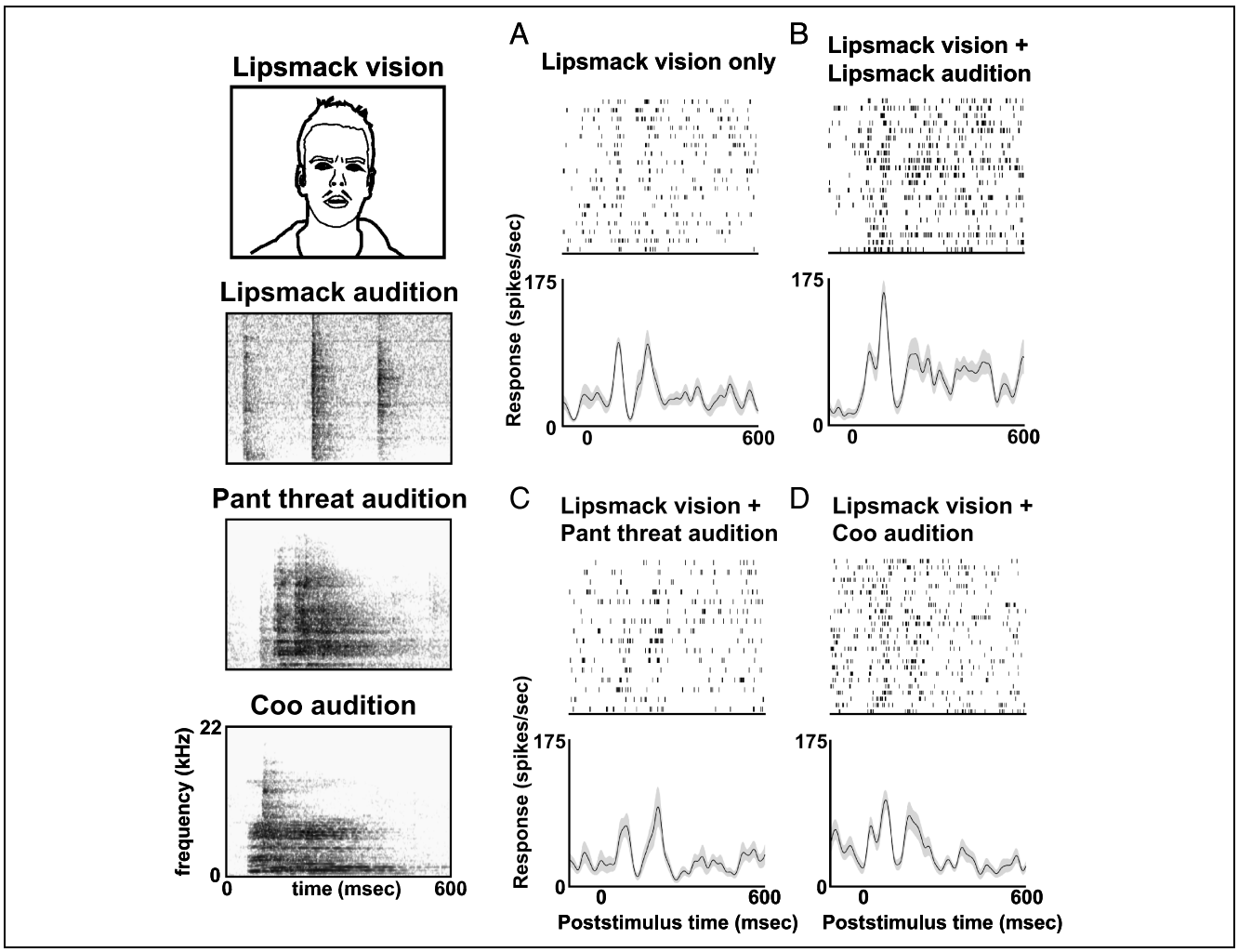

Figure 6. Augmentation of visual responses by congruent auditory stimuli. Top left: An illustration of the middle frame from a movie of a human face lip-smacking. Bottom left: Spectrograms of the different auditory stimuli, *x*-axis = time (msec) and *y*-axis = frequency (kHz), and amplitude represented by the darkness of the gray scale (white 50 dB, black 100 dB). Right: Single-cell responses to a movie of a human face lip-smacking presented with (A) the visual stimulus alone (trials = 18), (B) with visual and congruent lip-smacking auditory stimulus (trials = 27), (C) with visual stimulus and a macaque pant-threat auditory stimulus (trials = 16), and (D) with visual stimulus and a macaque coo auditory stimulus (trials = 25). The upper section of each plot shows individual trial responses as rastograms, the lower section the SDFs calculated from all trials (gray = *SEM*). Post hoc testing showed that the visual response to the human face lip-smacking was augmented significantly with the addition of the lip-smacking auditory stimulus ($p < .05$). This augmentation was not seen with the addition of either the pant-threat or the coo auditory stimulus ($p > .05$ each comparison).

animals. We saw no apparent concentration of cells showing audiovisual integration to one cortical region. Figure 8 shows the position of neurons from Monkey 1 tested for an interaction between the auditory and visual responses.

DISCUSSION

We find that 23% of STS neurons visually responsive to actions are modulated significantly by the corresponding auditory stimulus, much more than would have been found by chance. The addition of an auditory signal results in either an augmentation or an attenuation of the visual response in equal numbers of modulated neurons. The modulation of the visual response by auditory signals is substantial; an 86% increase in the average visual response of cells where auditory signals augment the visual responses and a 46% decrease in the average visual response of cells where auditory signals attenuate the visual response. Furthermore, for response augmentation, the integration is dependent upon the auditory stimulus matching the visually presented action.

We have found neurons showing audiovisual integration in the upper bank and fundus of the STS, previously defined as the polysensory region of the STS by Bruce et al. (1981), and in the lower bank also known to contain polysensory neurons (Benevento et al., 1977). These results are also consistent with reports of auditory

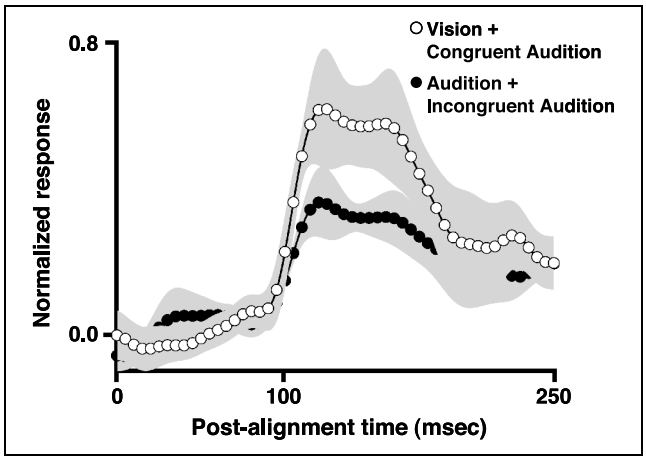

Figure 7. Responses to congruent and incongruent audiovisual combinations averaged across cells. SDFs (gray = *SEM*) from the responses of seven neurons in which auditory signals augmented the visual response that were tested with congruent and incongruent auditory stimuli. The response to the V condition is not illustrated because of the extensive overlap with the response to the combined visual and incongruent auditory stimulus.

stimuli (Gibson & Maunsell, 1997) and other modality stimuli (Colombo & Gross, 1994) modulating visually induced neural activity in regions of the temporal cortex outside of the upper bank of the STS. Both auditory and visual information are available to all regions within rostral STS either directly or indirectly from neighboring regions of cortex (Saleem et al., 2000; Seltzer & Pandya, 1978, 1984, 1989, 1994; see also Introduction).

Benevento et al. (1977) recorded from neurons in the upper bank and lower bank of the STS and found that a higher proportion (36%) of their recorded neurons were responsive to visual and auditory stimuli than we did here. They used visual stimuli consisting of moving bars and objects and flashes of light; auditory stimuli consisted of tones and clicks ranging from 100 Hz to 15 kHz. The potential discrepancy in the number of audiovisual neurons (36%, cf. our 23%) can be explained by differences in the range of stimuli used and the technique used to search for responses to both visual and auditory stimuli.

We wanted to know how neurons responding to visually presented actions would integrate sounds associated with those actions. Bruce et al. (1981) used stimuli perhaps more similar to ours. They tested the visual response properties of STS neurons in the upper bank and fundus with complex visual stimuli, including images of faces. Although they did not systematically test all neurons for auditory responses, they found 38% of the visually responsive neurons would also respond to the auditory stimuli (clicks, tones, jangling keys, and monkey calls). Bruce et al. also reported that "a few neurons" responded to the combined sight and sound of an object striking a surface but not one modality alone. Although we found three cells responsive to hitting actions modulated by the sound of that action, our screening phase for visually responsive cells would preclude us from finding neurons that code actions only when vision and sound are combined. Our findings of audiovisual integration in the STS substantiate prior studies of auditory and visual responses.

A measure of the degree of audiovisual integration is whether the combined response can be predicted by the addition of the two unimodal responses (Stein & Meredith, 1993). In the macaque's superior colliculus (Wallace et al., 1996) and the cat's anterior ectosylvian sulcus (Wallace, Meredith, & Stein, 1992), audiovisual integration can be indicated by a multimodal response greater than the sum of the responses to the individual inputs themselves (supra-additive). This effect can be seen in the BOLD response in human STS when the sight of a face speaking is combined with the sound of the speech (Calvert, 2000). Similar to previous physiology (Wallace et al., 1992, 1996), and in support of the human fMRI results (Calvert, 2000), our results indicate that STS neurons showing an augmentation of the visual response by the addition of auditory signals typically exhibited supra-additive integration.

For the majority of visual cells tested, auditory stimuli alone produce very little response in the cells (Figure 4). On the other hand, the effect of the auditory stimulus on the visual response is substantial and is present at response onset. Current source densities, an index of synaptic input, measured in the STS in response to simple auditory tones, have latencies of 25 msec (Schroeder & Foxe, 2002). Neurons responding to auditory stimuli have latencies of approximately 45 msec, where visual response latencies are later (e.g., 140 msec) (Bruce et al., 1981). Thus, early auditory inputs to the region are available to modulate the later visual response.

Benevento et al. (1977) noted the existence of one neuron's visual response to a moving bar being attenuated by a tone of constant frequency. This type of multisensory interaction is seen in the superior colliculus if visual and auditory stimuli are presented in different regions of space or desynchronized in time (Stein & Meredith, 1993). Examining all STS neurons that had an attenuation of the visual response by the addition of auditory signals, we found on average that the integration was sub-additive. This attenuation of the visual response is not dependent upon the type of auditory stimulus matching the visual stimulus as incongruent auditory stimuli have a similar effect to the congruent auditory stimuli on the visual response.

In neurons where the visual response is augmented by auditory signals, sounds incongruent to the visually presented action do not affect the visual response to the same extent as congruent sounds. Response augmentation (but not necessarily response attenuation) is therefore dependent upon the matching of the sound of an action to the sight of the action. Combining the visual stimulus of a speaking face with incongruent speech results in a sub-additive BOLD response in human STS

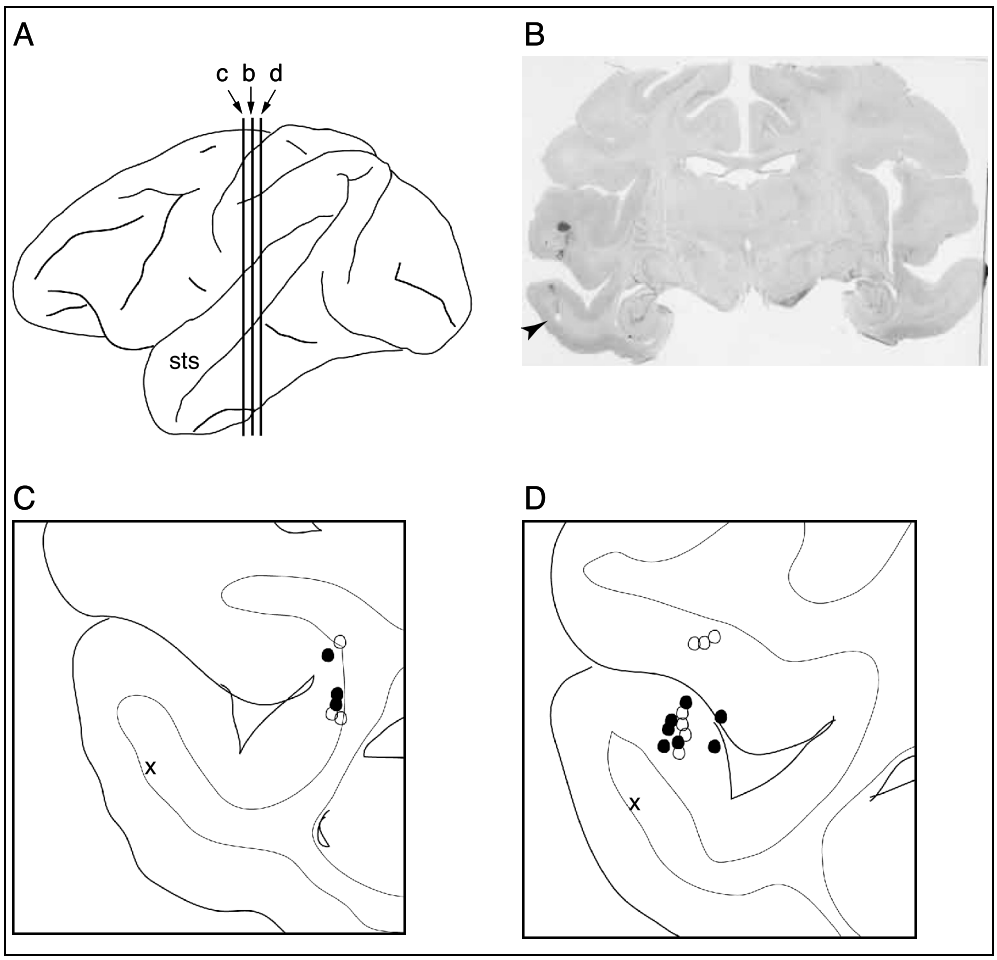

Figure 8. Histology and reconstruction of coronal sections in one monkey illustrating the position of neurons with audiovisual interactions. (A) Positions of the sections along the STS illustrated on a schematic representation of the brain, (B) +8.5, (C) +10, and (D) +7, anterior with respect to the interaural line. (B) Nissil-stained section with the position of an electrode lesion marked by the black arrow. (C,D) Two reconstructions enlarged to show the STS gray and white matter boundaries. The positions of the lesion (cross), neurons that integrate visual and auditory stimuli (solid circles) and neurons that do not integrate visual and auditory stimuli (open circles) are marked on each reconstruction.

(Calvert, 2000). fMRI measures of the BOLD signal reflect the responses of a population of neurons within each ''voxel.'' To make an appropriate comparison between the BOLD signal and the cell responses, we should take into account the responses of all cells showing audiovisual integration, as well as cells that respond to vision alone and those that respond to audition alone. When congruent visual and auditory stimuli are presented, unimodal visual and unimodal auditory cells will be active. Cells that have visual responses augmented with congruent stimuli will show response augmentation and cells that have visual responses attenuated with congruent stimuli will show response attenuation. Under conditions when visual stimuli are presented with incongruent auditory stimuli, the same unimodal visually responsive cells will be active and a different population of unimodal acoustically responsive cells will be active. Critically, cells that have visual responses augmented with congruent stimuli will have little or no response augmentation, whereas those cells that have visual responses attenuated with congruent stimuli will still show response attenuation. Thus, there will be reduced total neural activity to incongruent compared to congruent visual and auditory stimuli. Thus, the results from the present study suggest a more detailed description or the results from related fMRI studies.

How can we explain why a proportion of STS neurons' visual responses attenuate with the addition of auditory signals? One explanation might be that an STS neuron's selectivity to a visual stimulus is difficult to explore fully because that would imply that all possible visual stimuli were tested. Although in our screening phase and experimental test sets we included several different actions, it is quite possible that this did not include the action that would have elicited the biggest response. Therefore, when the auditory stimulus was presented with the visual stimulus, this information could ''confirm'' that the combined stimulus was not the ''correct'' action for a cell under testing and hence we would see a visual response attenuation. In our population of neurons with visual response augmentation by auditory signals, however, incongruent stimuli did not produce response attenuation, as would be predicted. A second possibility is that receptive field misalignment in the attenuated neurons might be a contributing factor to their depressed responses, similar to the effects seen in some superior colliculus neurons (Wallace et al., 1996). This explanation, however, cannot account for why we see augmentation of visual responses with the same

stimuli in other neurons, and why we see, for 6 of 14 (43%) of cells showing attenuation, an auditory augmentation of a different visual response. Additionally, most neurons recorded in the STS have visual and auditory receptive fields that extend over a larger region of space than covered by the stimuli used in this study (Hikosaka et al., 1988; Bruce et al., 1981). Therefore, we suggest that neurons that have visual response attenuation by auditory signals might represent a separate population of cells whose functional role is as yet undetermined.

In humans, matching of two unimodal stimuli increases the likelihood and speed of a response to the combined cross-modal stimulus over the response to the unimodal stimulus (Fort, Delpuech, Pernier, & Giard, 2002; Tipper et al., 2001; Giard & Peronnet, 1999; Driver, 1996). Increased activity seen here in STS neurons where the visual response is augmented by the addition of an auditory signal could lie behind an increased likelihood of detecting the cross-modal stimulus. The STS is involved in processing higher order statistics of the visual and auditory stimuli. The audiovisual integration shown here in the responses of these STS neurons represents multisensory integration at a single-cell level. Response enhancement in the population of neurons that have visual responses augmented by combining congruent visual and auditory stimuli could help those neurons code actions and enhance their detection. The correct matching of sight and sound may result in enhanced action detection due to a higher neuronal response rate; incorrect matching with a relatively lower neuronal response rate may result in relatively reduced action detection.

STS neurons appear to integrate the sight of facial actions, hand actions, and body movements with sounds arising from these actions and interaction with the environment. Output from these neurons may help resolve conflicting social signals (Ghazanfar & Logothetis, 2003) and provide a better representation of the meaning of the social signal. In the superior colliculus, sights and sounds are matched on spatial location and time of occurrence (Wallace et al., 1996; Meredith & Stein, 1986a, 1996). Matching of sights and sounds by multisensory superior collicular neurons appears to lie behind multisensory orientation responses (Burnett, Stein, Chaponis, & Wallace, 2004; Jiang, Jiang, & Stein, 2002). Both the integration in cat's superior colliculus and associated orienting behaviors are dependent upon cortical inputs (Jiang & Stein, 2003; Jiang et al., 2002; Jiang, Wallace, Jiang, Vaughan, & Stein, 2001). The STS in the monkey has substantial reciprocal connections with the superior colliculus by way of the pulvinar (Lui, Gregory, Blanks, & Giolli, 1995; Burton & Jones, 1976; Benevento & Fallon, 1975). The extent to which multisensory integration in primate STS and superior colliculus is reliant upon each other, however, is yet to be determined. STS neurons are often sensitive to the highly complex visual stimuli of actions, such as hand object interactions and other socially important signals such as facial movement. Thus, in this study, we used action stimuli (the most effective tested stimuli) to test the effect of sound on the visual response. It will be important to know how space, time, acoustic properties, and functional referents determine STS integration to understand how STS integration relates to multimodal integration seen in the primate superior colliculus. Principle component analysis and independent component analysis have been recently used to examine the sensitivity of neurons in the ventral lateral prefrontal cortex to acoustic features of macaque vocalizations (Averbeck & Romanski, 2004) and could be gainfully used to determine the different factors behind multisensory integration in the primate STS.

Responses to observed motor acts have shorter response latencies in temporal cortex than in the premotor cortex (Nishitani & Hari, 2000). To our knowledge, there has been no evidence of STS polysensory neurons responding during the execution of motor acts; active search for such properties has not been successful (Christian Keysers, personal communication). We propose that STS neurons form a multisensory representation of actions without any necessary reference to motor circuits. The integration of auditory and visual representations at the level of the STS could be passed on directly, or indirectly via the parietal cortex, to the premotor cortex.

METHODS

Physiological Subjects, Recording, and Reconstruction Techniques

Two rhesus macaques, aged 6 and 9 years, were trained to sit in a primate chair with head restraint. Using standard techniques (Perrett et al., 1985), recording chambers were implanted over both hemispheres to enable electrode penetrations to reach the STS. Single neurons were recorded using tungsten microelectrodes inserted through the dura. The subject's eye position ($\pm 1°$) was monitored (IView, SMI, Germany). A Pentium IV PC with a Cambridge electronics CED 1401 interface running Spike 2 recorded eye position, spike arrival, and stimulus on/offset times.

After each electrode penetration, X-ray photographs were taken coronally and parasagitally. The positions of the tip of each electrode and its trajectory were measured with respect to the intra-aural plane and the skull's midline. Using the distance of each recorded neuron along the penetration, a three-dimensional map of the position of the recorded cells was calculated. Coronal sections were taken at 1-mm intervals over the anterior–posterior extent of the recorded neurons. Alignment of sections with the X-ray coordinates of the recording sites was achieved using the location of microlesions and injection markers on the sections.

Stimuli

Stimuli consisted of 24-bit color pictures of objects and body parts or short (360–2352 msec) 16-bit color movies of humans walking, hand actions, and facial movements (actions) performed by the human experimenter (N.E.B.) or monkeys from the home colony. This ensured that familiar individuals performed all actions. Actions were filmed with a 3CCD digital video camera (Panasonic, NV-DX110), against a plain background. Actions were classified into those that could normally be associated with a sound (e.g., tearing paper) and those actions that were normally silent (grasping an object, grooming fur, raising, lowering, and rotating the head). Only actions that were normally associated with a sound were used in the experimental phase of the study. Walking actions consisting of a human walking to the left or right, toward or away, showing the whole body. Hand (and foot) actions, performed by a human actor, included tearing a piece of paper with two hands, manipulating a piece of paper with one hand (also home colony macaque actor), striking a wall with a fist, hitting a desk with the palm, and kicking a wall with one leg. Facial actions consisted of a shout, lip-smack, chew (human), pant-threat, and coo (macaque–cage mate). Macaque vocalizations were defined by the context in which they were performed, a more accurate method than by acoustics as some (e.g., grunt and pant-threat), appear acoustically similar (Tecumseh Fitch, personal communication). The "coo" exemplar was performed by the cage mate calling to one subject after being removed from the room; the "pant-threat" exemplar was performed by the cage mate when threatening the experimenter.

In the experimental phase, actions were presented under different conditions: VA, V, A, VAi, and with the "incongruent" auditory stimulus alone (Ai). Congruency was established by recording the actual action with the digital video camera. Because a judgment of "incongruency" could be subjective, we tested with different incongruent auditory stimuli, including the original soundtrack time reversed and auditory stimuli from other recorded actions. In this article, all the responses to the different VAi stimuli are considered together and all the responses to the different Ai stimuli are considered together.

Stimulus Presentation

Visual stimuli were presented centrally on a black monitor screen (Sony GDM-20D11, resolution 25.7 pixels/deg, refresh rate 72 Hz), 57 cm from the subject. Auditory stimuli were presented through speakers (Samsung, HK395). Two speakers were positioned 57 cm from the subject, 30° left and right from the center of the image. The third subwoofer speaker was positioned below the monitor.

Screening stimuli were stored on an Indigo2 Silicon Graphics workstation hard disk. Static images subtending 19° × 19° were presented for 125 msec. Actions were presented by rendering a series of bitmaps subtending 19° × 19° or 25° × 20.5°; each bitmap in the series was presented for 42 msec and represented one frame of the movie. The number of frames per bitmap in each action ranged from 9 to 56 per action.

Experimental stimuli were stored on VHS videotape. Auditory information was recorded onto one audio channel, the other audio channel was reserved for signaling the onset and offset of stimulus presentation to the PC. Different actions lasted from 360 to 2320 msec, and the interstimulus interval was 250 ± 40 msec. Each stimulus was recorded 48 times in a pseudorandom order, with the constraint that no stimulus was presented for the $n + 1$-th time until each stimulus was presented n times. VHS tapes were played back through the Silicon Graphics workstation under control of the PC running Spike2 and recording spikes. Visual stimuli subtended 25° × 20.5° at 57 cm, and the auditory stimuli were played through the three speakers at a sound pressure level of 65–85 dB. Buffering of the visual image introduced a visual stimulus lag relative to the auditory stimulus of 14–28 msec.

Testing Procedure

Cell responses were isolated using standard techniques, and visualized using an oscilloscope. Preliminary clinical screening was performed with some STS neurons to assess the extent of response sensitivity to the sight and sound of actions. This was carried out using human actors clapping, speaking, foot tapping, and walking. These actions were performed in and out of sight to measure the response to the sound of the action and the sight and sound of the action. More systematic screening was performed with a search set of (on average 55) images and movies of different objects, body parts, and actions previously shown to activate neurons in the STS (Foldiak, Xiao, Keysers, Edwards, & Perrett, 2003). Static images and silent movies were presented centrally on the monitor in a random sequence with a 500-msec interstimulus interval. Presentation of this screening set commenced when the subject fixated (±3°) a yellow dot presented centrally on the screen for 500 msec (to allow for blinking, deviations outside the fixation window lasting <100 msec were ignored). Fixation was rewarded with the delivery of fruit juice. Spikes were recorded during the period of fixation, if the subject looked away for longer than 100 msec, spike recording and presentation of stimuli stopped until the subject resumed fixation for >500 msec. Responses to each stimulus in the screening set were displayed as on-line rastergrams and poststimulus time histograms (PSTHs) aligned to stimulus onset. Cells that showed a clear visual response

to an action in the screening set associated with sound were then tested in the experimental phase.

In the experimental phase, each cell was tested with a stimulus set containing actions in which each action was represented under the VA, V, and A conditions. The stimulus set presented contained the action that showed a clear visual response in the screening phase of the experiment and other actions. One stimulus set contained 12 different actions; other stimulus sets contained two to five similar actions grouped by action type (e.g., walking, hand actions, or face actions). This procedure was performed to allow us to reconfirm statistically (see below) the action type that showed the biggest visual response as determined earlier in the screening phase and to measure audiovisual integration for this action. Additionally, this allowed us to test whether sound affected the visual responses to less effective actions. Initially in this study, neurons were tested with the stimulus set containing 12 actions represented under the VA, V, and A conditions. Having found that sound modulated the visual response of STS neurons, we then subsequently tested all neurons exclusively with stimulus sets containing actions presented under the VA, V, A, VAi, and Ai conditions for auditory–visual interactions. This then allowed us to answer the additional question of whether the type of sound mattered. Responses were recorded for the duration of the stimulus tape (48 trials per condition) or until the neural response was lost, whichever was earlier, while the subject fixated the yellow dot (±3°). Neural signals were recorded to hard disk for off-line filtering and analysis.

Response Analysis

Single-Cell Analysis

Off-line isolation of single cells was performed using a template-matching procedure and principal components analysis (Spike2). Each cell's response to a stimulus in the experimental test set was calculated by aligning segments in the continuous recording, on each occurrence of that particular stimulus (trials). Eye movement information was used to include only those trials where the subject was fixating for over 80% of the first 300 msec of stimulus presentation.

For each stimulus, a PSTH was generated and an SDF calculated by summing across trials (bin size = 1 msec) and smoothing (Gaussian, σ = 10 msec). Background firing rate was measured in the 100-msec period before stimulus onset. Response latencies to each stimulus were measured as the first 1-msec time bin where the SDF exceeded 3 standard deviations above the background firing rate for over 15 msec in the 0- to 300-msec period after stimulus onset (Edwards, Xiao, Keysers, Foldiak, & Perrett, 2003; Oram, Perrett, & Hietanen, 1993; Oram & Perrett, 1992, 1996).

Responses to each action under the different conditions were measured within a 60-msec window starting at the average VA and V cell response latency. For each cell, the responses to an action under the three different conditions were entered into a two-way ANOVA [action (n = 2–12) by condition (V, A, VA) with trials as replicates], if there was a significant visual response to the action and all the conditions contained more than five trials.

Two types of result from the ANOVA indicated that the auditory signal may have had a significant effect on the visual response of the cell. A significant main effect of condition ($p < .05$) indicated a possible difference in the visual responses with and without auditory signals. Post hoc analysis (PLSD) was used to determine the significance of the difference between the responses to the VA and V conditions. The responses to the action that elicited the biggest combined response to all conditions were taken in subsequent population analyses. Second, significant interaction between action and condition ($p < .05$) also indicated that for the cell auditory signals might significantly affect the visual responses to some actions. The PLSD value was used to test whether there was a difference between VA and V responses to each action separately. A significant difference between the VA and V conditions for one of the actions indicated that the visual response to that action was augmented (VA > V) or attenuated (VA < V) by auditory signals. For each action that showed a significant effect of sound, the responses to the VA, V, and A conditions were summed to calculate the size of the response to the action under all the different conditions. The responses to the action with the largest "combined" response from the actions with VA > V, and the responses to the action with the largest "combined" responses from the actions with VA < V were taken in subsequent population analyses.

For those cells showing a statistically significant modulatory effect of sound on the visual response, the following formula was used to calculate the size of the change in the visual response with the addition of auditory signals.

$$Magnitude\,(\%) = \frac{VA - V}{V} \times 100$$

where VA is the response to the visual and auditory stimuli combined and V is the response to the visual stimulus presented alone, all responses were measured after subtracting the mean background firing rate.

The following formula below was used to calculate an index of the linearity of the integration of the visual and auditory signals:

$$I_{audvis} = \frac{VA - (V + A)}{V + A}$$

where VA and V are as above and A is the response to the auditory stimulus alone, all responses were mea-

sured after subtracting the mean background firing rate. A value of zero for I_{audvis} indicates that the response to the combined audiovisual stimulus can be predicted by a simple addition of the responses to the visual stimulus presented alone and the auditory stimulus presented alone. Values greater than zero indicate that the response to the combined audiovisual stimulus is greater than the sum of the responses to each modality, a nonlinear supra-additive integration. Values less than zero indicate that the response to the combined audiovisual stimulus is less than the sum of the responses to each modality, a nonlinear sub-additive integration.

Population Analysis

Responses of cells whose visual response was augmented by the addition of an auditory signal (VA > V) and responses of cells whose visual response was attenuated by an auditory signal (VA < V) were combined separately to create an average cell with visual response augmentation and an average cell with visual response attenuation. First, for the average cell with a visual response augmentation, each contributing cell's SDF in the VA, V, and A conditions were normalized with respect to the peak response in the VA condition. This was done so that every cell included contributed equally to the population response. Second, each cell's SDF to the V condition was shifted in time such that the cell's visual response latency was aligned at 100 msec. The corresponding SDF for the VA and A conditions were also shifted an equivalent amount. To obtain the responses of an average cell with visual response attenuation, the responses were calculated in a similar manner as described above, except the responses were normalized to the peak response in the V condition.

Acknowledgments

This work was supported by grants from the European Union and the Wellcome Trust.

Reprint request should be sent to David Perrett, School of Psychology, St Mary's College, University of St Andrews, South Street, St Andrews, Fife KY16 9JP, Scotland, UK, or via e-mail: dp@st-andrews.ac.uk.

REFERENCES

Averbeck, B. B., & Romanski, L. M. (2004). Principle and independent components of macaque vocalizations: Constructing stimuli to probe high-level sensory processing. *Journal of Neurophysiology, 91,* 2897–2909.

Baylis, G. C., Rolls, E. T., & Leonard, C. M. (1987). Functional subdivisions of the temporal neocortex. *Journal of Neuroscience, 7,* 330–342.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron, 41,* 809–823.

Bekkering, H., & Wohlschlaeger, A. (2002). Action perception and imitation: A tutorial. In W. Prinz, & B. Hommel (Eds.), *Attention and performance: XIX. Common mechanisms in perception and action* (pp. 294–333). Oxford: Oxford University Press.

Bell, A. H., Corneil, B. D., Munoz, D. P., & Meredith, M. A. (2003). Engagement of visual fixation suppresses sensory responsiveness and multisensory integration in the primate superior colliculus. *European Journal of Neuroscience, 18,* 2867–2873.

Benevento, L. A., & Fallon, J. H. (1975). The ascending projections of the superior colliculus in the rhesus monkey (*Macaca mulatta*). *Journal of Comparative Neurology, 160,* 339–362.

Benevento, L. A., Fallon, J., Davis, B. J., & Rezak, M. (1977). Auditory–visual interaction in single cells in the cortex of the superior temporal sulcus and the orbital frontal cortex of the macaque monkey. *Experimental Neurology, 57,* 849–872.

Bruce, C. J., Desimone, R., & Gross, C. G. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of Neurophysiology, 46,* 369–384.

Bruce, C. J., Desimone, R., & Gross, C. G. (1986). Both striate cortex and superior colliculus contribute to visual properties of neurons in superior temporal polysensory area of macaque monkey. *Journal of Neurophysiology, 55,* 1057–1075.

Burnett, L. R., Stein, B. E., Chaponis, D., & Wallace, M. T. (2004). Superior colliculus lesions preferentially disrupt multisensory orientation. *Neuroscience, 124,* 535–547.

Burton, H., & Jones, E. G. (1976). The posterior thalamic region and its cortical projection in New World and Old World monkeys. *Journal of Comparative Neurology, 168,* 249–302.

Bushara, K. O., Hanakawa, T., Immisch, I., Toma, K., Kansaku, K., & Hallett, M. (2003). Neural correlates of cross-modal binding. *Nature Neuroscience, 6,* 190–195.

Calvert, G. A. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology, 10,* 649–657.

Calvert, G. A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex, 11,* 1110–1123.

Colombo, M., & Gross, C. G. (1994). Responses of inferior temporal cortex and hippocampal neurons during delayed matching to sample in monkeys (*Macaca fascicularis*). *Behavioural Neuroscience, 108,* 443–455.

Desimone, R., & Gross, C. G. (1979). Visual areas in the temporal cortex of the macaque. *Brain Research, 178,* 363–380.

Desimone, R., & Ungerleider, L. G. (1986). Multiple visual areas in the caudal superior temporal sulcus of the macaque. *Journal of Comparative Neurology, 248,* 164–189.

di Pelligrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research, 91,* 176–180.

Distler, C., Boussaoud, D., Desimone, R., & Ungerleider, L. G. (1993). Cortical connections of inferior temporal area TEO in macaque monkeys. *Journal of Comparative Neurology, 334,* 125–150.

Dodd, B. (1977). The role of vision in the perception of speech. *Perception, 6,* 31–40.

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for the visual processing of the human body. *Science, 293,* 2470–2473.

Driver, J. (1996). Enhancement of selective listening by illusory

mislocation of speech sounds due to lip-reading. *Nature, 381,* 66–68.
Edwards, R., Xiao, D.-K., Keysers, C., Foldiak, P., & Perrett, D. I. (2003). Color sensitivity of cells responsive to complex stimuli in the temporal cortex. *Journal of Neurophysiology, 90,* 1245–1256.
Ferrari, P. F., Gallese, V., Rizzolatti, G., & Fogassi, L. (2003). Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. *European Journal of Neuroscience, 17,* 1703–1714.
Foldiak, P., Xiao, D.-K., Keysers, C., Edwards, R., & Perrett, D. I. (2003). Rapid serial visual presentation for the determination of neural selectivity in area STSa. *Progress in Brain Research, 144,* 107–116.
Fort, A., Delpuech, C., Pernier, J., & Giard, M. H. (2002). Dynamics of cortico-subcortical cross-modal operations involved in audio-visual object detection in humans. *Journal of Cognitive Neuroscience, 12,* 1031–1039.
Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain, 119,* 593–609.
Ghazanfar, A. A., & Logothetis, N. K. (2003). Facial expressions linked to monkey calls. *Nature, 423,* 937–938.
Giard, M. H., & Peronnet, F. (1999). Auditory–visual integration during multimodal object recognition in humans: A behavioural and electrophysiological study. *Journal of Cognitive Neuroscience, 11,* 473–490.
Gibson, J. R., & Maunsell, J. H. R. (1997). Sensory modality specificity of neural activity related to memory in visual cortex. *Journal of Neurophysiology, 78,* 1263–1275.
Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., & Blake, R. (2000). Brain areas involved in perception of biological motion. *Journal of Cognitive Neuroscience, 12,* 711–720.
Hasselmo, M. E., Rolls, E. T., & Baylis, G. C. (1989). The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behavioural Brain Research, 32,* 203–218.
Hikosaka, K., Iwai, E., Saito, H., & Tanaka, K. (1988). Polysensory properties of neurons in the anterior bank of the caudal superior temporal sulcus of the macaque monkey. *Journal of Neurophysiology, 60,* 1615–1637.
Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazzionta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science, 286,* 2526–2528.
Jay, M. F., & Sparks, D. L. (1984). Auditory receptive fields in primate superior colliculus shift with changes in eye position. *Nature, 309,* 345–347.
Jiang, W., Jiang, H., & Stein, B. E. (2002). Two corticotectal areas facilitate multisensory orientation behaviour. *Journal of Cognitive Neuroscience, 14,* 1240–1255.
Jiang, W., & Stein, B. E. (2003). Cortex controls multisensory depression in superior colliculus. *Journal of Neurophysiology, 90,* 2123–2135.
Jiang, W., Wallace, M. T., Jiang, H., Vaughan, J. W., & Stein, B. E. (2001). Two cortical areas mediate multisensory integration in superior colliculus neurons. *Journal of Neurophysiology, 85,* 506–522.
Kadunce, D. C., Vaughan, J. W., Wallace, M. T., Benedek, G., & Stein, B. E. (1997). Mechanisms of within- and cross-modality suppression in the superior colliculus. *Journal of Neurophysiology, 78,* 2834–2847.
Kohler, E., Keysers, C., Umilta, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science, 297,* 846–848.
Lui, F., Gregory, K. M., Blanks, R. H. I., & Giolli, R. A. (1995). Projections from visual areas of the cerebral cortex to pretectal nuclear complex, terminal accessory optic nuclei, and superior colliculus in macaque monkey. *Journal of Comparative Neurology, 363,* 439–460.
McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.
Meredith, M. A., & Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus. *Science, 221,* 389–391.
Meredith, M. A., & Stein, B. E. (1986a). Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Brain Research, 365,* 350–354.
Meredith, M. A., & Stein, B. E. (1986b). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of Neurophysiology, 56,* 640–662.
Meredith, M. A., & Stein, B. E. (1996). Spatial determinants of multisensory integration in cat superior colliculus. *Journal of Neurophysiology, 75,* 1843–1857.
Nishitani, N., & Hari, R. (2000). Temporal dynamics of cortical representation for action. *Proceedings of the National Academy of Sciences, U.S.A., 97,* 913–918.
Oram, M. W., & Perrett, D. I. (1992). Time course of neural responses discriminating views of the face and head. *Journal of Neurophysiology, 68,* 70–84.
Oram, M. W., & Perrett, D. I. (1994). Responses of anterior superior temporal polysensory (STPa) neurons to "biological motion" stimuli. *Journal of Cognitive Neuroscience, 6,* 99–116.
Oram, M. W., & Perrett, D. I. (1996). Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey. *Journal of Neurophysiology, 76,* 109–129.
Oram, M. W., Perrett, D. I., & Hietanen, J. K. (1993). Directional tuning of motion-sensitive cells in the anterior superior temporal polysensory area of the macaque. *Experimental Brain Research, 97,* 274–294.
Perrett, D. I., Harries, M. H., Bevan, R., Thomas, S., Benson, P. J., Mistlin, A. J., Chitty, A. J., Hietanen, J. K., & Ortega, J. E. (1989). Frameworks of analysis for the neural representation of animate objects and actions. *Journal of Experimental Biology, 146,* 87–113.
Perrett, D. I., Mistlin, A. J., Harries, M. H., & Chitty, A. J. (1990). Understanding the visual appearance and consequences of hand actions. In M. A. Goodale (Ed.), *Vision and action: The control of action* (pp. 163–180). Norwood, NJ: Ablex.
Perrett, D. I., Smith, P. A. J., Mistlin, A. J., Chitty, A. J., Head, A. S., Potter, D. D., Broennimann, R., Milner, A. D., & Jeeves, M. A. (1985). Visual analysis of body movements by neurons in the temporal cortex of the macaque monkey: A preliminary report. *Behavioural Brain Research, 16,* 153–170.
Perrett, D. I., Smith, P. A. J., Potter, D. D., Mistlin, A. J., Head, A. S., Milner, A. D., & Jeeves, M. A. (1984). Neurons responsive to faces in the temporal cortex: Studies of functional organization, sensitivity to identity and relation to perception. *Human Neurobiology, 3,* 197–208.
Prinz, W. (2002). Experimental approaches to immitation. In A. N. Melzoff & W. Prinz (Eds.), *The imitative mind. Development, evolution and brain bases* (pp. 143–162). Cambridge: Cambridge University Press.
Puce, A., Allison, T., Bentin, S., Gore, J. C., & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *Journal of Neuroscience, 18,* 2188–2199.
Puce, A., & Perrett, D. I. (2003). Electrophysiology and brain

imaging of biological motion. *Philosophical Transactions of the Royal Society B, 358,* 435–445.
Rauschecker, J. P., Tian, B., & Hauser, M. (1995). Processing of complex sounds in the macaque nonprimary auditory cortex. *Science, 268,* 111–114.
Saleem, K. S., Suzuki, W., Tanaka, K., & Hashikawa, T. (2000). Connections between anterior inferotemporal cortex and superior temporal sulcus regions in the macaque monkey. *Journal of Neuroscience, 20,* 5083–5101.
Schroeder, C. E., & Foxe, J. J. (2002). The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. *Cognitive Brain Research, 14,* 187–198.
Seltzer, B., & Pandya, D. N. (1978). Afferent cortical connections and architectonics of the superior temporal sulcus and surrounding cortex. *Brain Research, 149,* 1–24.
Seltzer, B., & Pandya, D. N. (1984). Further observations on parietal–temporal connections in the rhesus monkey. *Experimental Brain Research, 55,* 301–312.
Seltzer, B., & Pandya, D. N. (1989). Intrinsic connections and architectonics of the superior temporal sulcus in the rhesus monkey. *Journal of Comparative Neurology, 290,* 451–471.
Seltzer, B., & Pandya, D. N. (1994). Parietal, temporal and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: A retrograde tracer study. *Journal of Comparative Neurology, 15,* 445–463.
Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses.* Cambridge: MIT Press.
Sugase, Y., Yamane, S., Ueno, S., & Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature, 400,* 869–873.
Tipper, S. P., Phillips, N., Dancer, C., Lloyd, D., Howard, L. A., & McGlone, F. (2001). Vision influences tactile perception at body sites that cannot be viewed directly. *Experimental Brain Research, 139,* 160–167.
Ungerleider, L. G., & Desimone, R. (1986). Cotical connections of visual area MT in the macaque. *Journal of Comparative Neurology, 248,* 190–222.
Viviani, P. (2002). Motor competence in the perception of dynamic events: A tutorial. In W. Prinz & B. Hommel (Eds.), *Attention and performance XIX. Common mechanisms in perception and action* (pp. 406–442). Oxford: Oxford University Press.
Von Bonin, G., & Bailey, P. (1947). *The neocortex of macaca mullata.* Urbana, IL: University of Illinois Press.
Wallace, M. T., Meredith, M. A., & Stein, B. E. (1992). Integration of multiple sensory modalities in cat cortex. *Experimental Brain Research, 91,* 484–488.
Wallace, M. T., Wilkinson, L. K., & Stein, B. E. (1996). Representation and integration of multiple sensory inputs in primate superior colliculus. *Journal of Neurophysiology, 76,* 1246–1266.