Exp Brain Res (2003) 153:628–636
DOI 10.1007/s00221-003-1603-5

REVIEW

C. Keysers · E. Kohler · M. A. Umiltà · L. Nanetti · L. Fogassi · V. Gallese

Audiovisual mirror neurons and action recognition

Received: 25 October 2002 / Accepted: 14 May 2003 / Published online: 23 August 2003

Abstract Many object-related actions can be recognized both by their sound and by their vision. Here we describe a population of neurons in the ventral premotor cortex of the monkey that discharge both when the animal performs a specific action and when it hears or sees the same action performed by another individual. These 'audiovisual mirror neurons' therefore represent actions independently of whether these actions are *performed, heard* or *seen.* The magnitude of auditory and visual responses did not differ significantly in half the neurons. A neurometric analysis revealed that based on the response of these neurons, two actions could be discriminated with 97% accuracy.

Keywords Audiovisual mirror neurons · Action recognition · Object-related actions · Ventral premotor cortex · Monkey · Neurometric analysis

Introduction

Understanding what someone else is doing is a process that is independent of the modality through which we perceive his actions: whether we hear or see someone knocking on our door makes no difference—we intuitively feel that knocking on the door is the same thing whether heard or seen. Indeed, we also intuitively grasp that knocking is the same when we do it ourselves, and when other people do it. While these statements seem trivial, understanding what brain mechanisms reside behind the brain's capacity to extract a single meaning—'knocking'—from such different modalities is far from trivial.

The rostral ventral premotor cortex (area F5, Fig. 1A) of the monkey contains a class of neurons called 'audiovisual mirror neurons' (Kohler et al. 2002) that might shed light on this question. By definition, 'mirror neurons' discharge both when a monkey makes a specific action and when it observes another individual making a similar action (Gallese et al. 1996; Rizzolatti et al. 1996). Effective actions for mirror neurons are those in which a hand or mouth interacts with an object. Grasping or tearing apart objects are examples of such effective actions. About half of these neurons also respond when the final part of the observed action, critical in producing a response in full vision, is occluded from sight (Umiltà et al. 2001) and can therefore only be 'guessed' by the monkey.

Recently (Kohler et al. 2002), we reported that a population of neurons called audiovisual mirror neurons additionally responds even when only the sound of the effective action is presented to the monkey. Non-hand-action related arousing sounds such as white noise or monkey vocalizations typically do not evoke significant responses in these neurons. These neurons respond differentially to different actions, and 22 of the 33 tested neurons responded more to a given action than to another independently of whether the actions were heard, seen or executed. Figure 1C, D illustrates these results.

The combination of motor, visual and auditory properties in these cells led us to hypothesize that audiovisual mirror neurons may be part of a network of neurons underlying our ability to discriminate actions independently of whether they are heard, seen or executed. This hypothesis raised two questions that will be addressed in the present paper. First: how do the visual and the

The first two authors contributed equally to the work

C. Keysers (✉) · E. Kohler · M. A. Umiltà · V. Gallese (✉)
Department of Neuroscience,
Università di Parma,
Via Volturno 39, 43100 Parma, Italy
e-mail: keysers@unipr.it
e-mail: vittorio.gallese@unipr.it

L. Nanetti
Department of Mathematics,
Università di Ferrara,
Via Machiavelli, 44100 Ferrara, Italy

L. Fogassi
Department of Psychology,
Università di Parma,
Bgo Carissimi 10, 43100 Parma, Italy

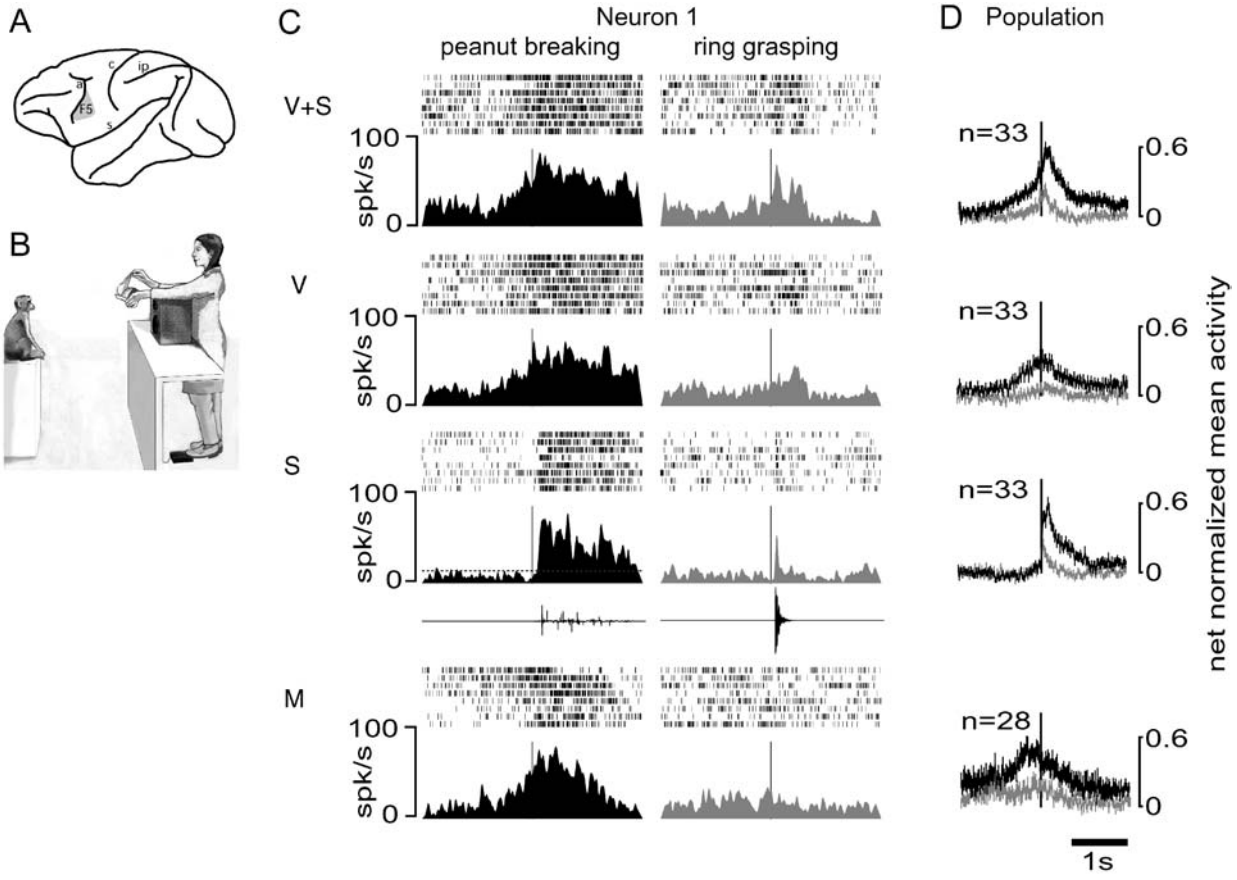

Fig. 1 **A** Lateral view of the macaque brain with the location of area F5 shaded in *gray*. Major sulci: *a* arcuate, *c* central, *ip* intraparietal, *s* sylvian sulcus. **B** Experimental setup; see "Materials and methods" for details. **C** Response of a neuron (Neuron 1) discriminating between two actions in all modalities. Rastergrams are shown together with spike density functions for the best (*black*) and the less effective action (*gray*). *V+S, V, S* and *M* stand for Vision-and-Sound, Vision-only, Sound-only and Motor conditions, respectively. The *vertical lines* indicate the time at which the sound occurred (*V+S, S*) or would have occurred (*V*).The *traces* under the spike-density functions in the sound-only conditions are oscillograms of the sounds played back to test the neurons. This neuron discharged when the monkey broke a peanut (*row M*) and when the monkey observed the experimenter making the same action (*rows V* and *V+S*). The same neuron also responded when the monkey only heard the sound of a peanut being broken without seeing the action (*row S*). When the monkey grasped a ring (*M*), Neuron 1 responded much less, demonstrating the motor specificity of the neuron. Also both the vision and the sound of an experimenter grasping the ring determined much smaller responses. A statistical criterion yielded both auditory and visual selectivity for this neuron. Note that in the S condition there is nothing for the monkey to see or hear prior to the onset of the action sound, and the neuron therefore remained silent prior to the onset of the action sound (*vertical line*). These trials therefore enable us to measure the auditory response onset latency of the neuron as the moment at which the activity after the sound onset goes above the mean ± 1.96 SD of the spontaneous activity prior to the sound (*dotted horizontal line*). In contrast, in the V+S and V conditions, the monkey sees preparatory parts of the action before the sound onset (e.g., the experimenter reaches for the peanut that he will later break), and the activity progressively increased prior to the sound onset, being elevated during the entire 2 s prior to the sound onset compared with the S condition. The same is true when the monkey himself performs the action (*M*). A similar effect was seen in most neurons. **D** Mean (± SEM) responses of the population of 33 tested neurons as a function of time relative to the auditory response onset latency (*vertical line*). The action that produced the strongest response when tested in vision and sound (best action, shown in *black*) determined stronger responses compared with the less effective action (*gray*) in all conditions. As for Neuron 1, the population maintained its action selectivity between modalities: the same action was more effective be it heard, seen or executed. (Adapted from Kohler et al. 2002)

auditory modalities interact to produce responses in audiovisual mirror neurons? Second: if audiovisual mirror neurons participate in our capacity to discriminate between actions independently of the modality through which they are perceived, can their firing reliably discriminate between different actions in all modalities? Using a neurometric analysis known as the Receiver Operator Characteristics (ROC) (Fig. 2) analysis, we therefore ask how well two actions could be discriminated based on the firing of audiovisual mirror neurons.

Materials and methods

Experimental animals and physiological procedures

These procedures have been explained elsewhere (Kohler et al. 2002). Briefly, three adult monkeys (*Macaca nemestrina*) were

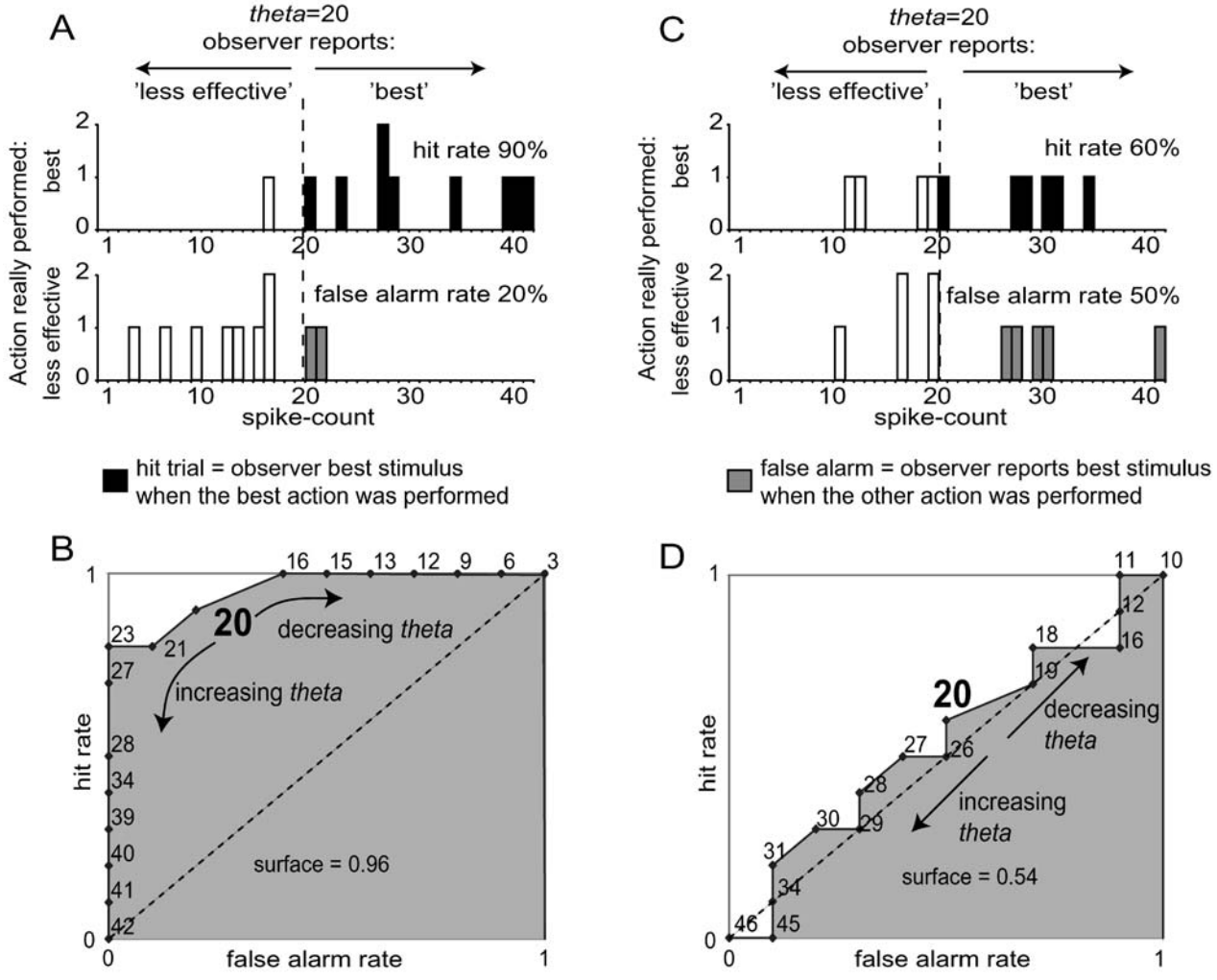

Fig. 2 Illustration of the ROC analysis for a neuron with a firing rate that does **(A, B)** and one that does not **(C, D)** discriminate between the two actions. **A** The spike count histogram of an audiovisual neuron is shown when the experimenter performed the best action in front of the monkey (*top histogram*) or the less effective action (*bottom*). The histogram illustrates the spike-count on the *x*-axis and the absolute frequency of observing this spike count on the *y*-axis. Note how the top histogram is shifted towards the right compared to the bottom histogram. This shift reflects the fact that high spike-counts are more likely if the best action was performed in front of the monkey. The *dotted vertical line* represents a *'theta'* value of 20 used by the observer to take its decisions: he reports the best action for spike counts larger than or equal to this value (*right of the dotted line*) and the less effective action for spike counts smaller than this value (*left of the dotted line*). Since the observer does not know which action the experimenter really performed in front of the monkey, he treats the two histograms equally. Responding with best action is correct (hit) when the best action was performed in front of the monkey (*black bars*) but a mistake (false alarm) when the less effective action was actually performed (*gray bars*). For this particular neuron and *theta*, there were nine hit trials in the ten best action trials (hit rate = 9/10=90%) and two false alarms in the ten less effective action trials (false alarm rate = 2/10=20%). **B** Shows the Receiver Operator Characteristic (ROC) curve for the neuron shown in **A**. This curve is obtained by linking the false alarm rate and hit rate obtained for all of the possible theta values. Each possible *theta* value is shown next to its respective point. **A** illustrates how to calculate the two values for a *theta* of 20 (shown as a *bold number*). Shifting the dotted *theta* line in **A** rightwards (i.e., increasing the *theta* value) mainly decreases the hit rate, resulting in a sharp dip of the curve left of the *theta*=20 point. If we move the *dotted line of* **A** leftwards, decreasing *theta*, we mainly increase the false alarm rate, resulting in the horizontal arm of the ROC curve rightwards of the *theta*=20 point. At all *theta*s, the hit rate remains above the false alarm rate for this neuron, resulting in an ROC curve that remains above the dotted diagonal. The surface under the ROC curve is 0.96. Such large surface values (close to 1) are typical for neurons that accurately discriminate between two actions. **C** Spike-count histograms for a neuron not discriminating between the two actions. Note how the two histograms largely overlap. For this neuron the terms best and less effective action are meaningless, and are only used in analogy to **A** and **B**. At *theta*=20, the observer has a hit rate of 60% and a false alarm rate of 50%. **D** ROC curve for the neuron of **C**. Increasing *theta* reduces hit and false alarm rate equally. Decreasing *theta* increases both values similarly. The curve remains along the diagonal, with a surface of 0.54. This behavior is typical for neurons not related to the observer's decision, with very overlapping histograms

trained to sit in a primate chair, head fixated and fitted with recording chambers. They performed hand actions on command for fruit juice reward. All experimental protocols were approved by the Veterinarian Animal Care and Use Committee of the University of Parma, and complied with the European law on the humane care and use of laboratory animals. Single neurons were recorded using tungsten microelectrodes (impedance: 0.5–1.5 MΩ measured at 1 kHz) inserted through the dura. Recording sites were attributed to area F5 based on topographical and physiological properties.

Data acquisition

The study described in the present article was preceded by an initial study in which, after discovering that some mirror neurons responded to auditory stimuli, we assessed whether these responses were due to arousal or other unspecific factors (Kohler et al. 2002). Since arousing control sounds did not evoke responses in F5 neurons, in the present study such control sounds were only tested occasionally and these results are not discussed in the present paper.

Whenever a neuron was isolated in area F5, general motor and visual properties were tested, as described previously (Gallese et al. 1996). All neurons were additionally tested with a battery of six actions that produced sounds ('noisy actions'): peanut breaking, ripping a sheet of paper, shaking a sheet of paper, crumpling a plastic bag, dropping a stick onto the floor, and grasping a metallic ring that emitted a sound when touched. Prior to recording, the monkeys were trained to perform all these actions on command: the target of the action was presented to the monkey, and fruit juice was given if the monkey performed the action. During training, the monkeys performed the natural version of the actions and the monkeys were therefore well acquainted with the acoustic consequences of all these actions. These six actions were tested a small number of times in front of the monkey. The most effective of these actions and one of the less effective actions at triggering a response in the neuron were then selected for further testing. To be fully tested using the paradigm described in the next paragraph, neurons had to respond to the sound of the best action. Responding to the sound of the best action was assessed by playing back the sound of the best action through a loudspeaker, and visually inspecting histograms of the response induced by this sound.

Full testing of the best and less effective actions then involved three 'sensory' conditions: vision-and-sound ('V+S'), vision-only ('V') and sound-only ('S') (see Fig. 1B), and during the active performance of the best, and in part of the neurons, the less effective action ('M'); see below. To test separately the visual and auditory contributions to the neuron's responses, the objects on which the actions were performed were modified so as to render the actions visually similar to the natural ones but silent. These silent versions of the action were: breaking an already broken peanut, ripping wet paper, shaking a sheet of rubber foam, crumpling rubber foam, dropping a stick onto a sheet of rubber foam. In the case of grasping the metal ring, the metallic sound was not played back through the loudspeaker. The absence of sound during silent actions was controlled using a sound level meter (Lutron SL-4001).

In all cases, pressing a foot pedal hidden from the monkey's sight triggered the recording of 4 s of spikes centered upon the pedal-pressing event. In V and V+S conditions pedal pressing was done just before the moment at which the action-sound normally starts. In S and V+S conditions, pedal pressing additionally triggered the playback of the pre-recorded sounds. By pressing the foot pedal at the adequate point in time, the experimenter synchronized the playback of the action-sound with the vision of the action. This gave the impression of a natural action. Although experimenters were very good at synchronizing the moment at which the action was performed with the pedal-pressing, there might still be a trial-to-trial variability in the order of ±20 ms in the synchronization with the visual stimulus. Synchronization with the auditory stimulus is within ±1 ms, due to an automatic triggering of the playback. Given that most analyses in this paper are done within windows of analysis over 1 s in length, this small jitter has no significant effect on the results.

In 14 of the cells, *x* and *y* eye position as measured using an infrared oculometer with a resolution of 1–5 min arc (Dr. Bouis, Germany; see Bach et al. 1983 for further details) were recorded in addition to spike activity. Statistical analysis of eye movements revealed that eye movements did not explain a significant proportion of the variance of the firing rates between conditions and they are therefore not discussed in the paper (all $p>0.05$).

Playback

Acoustic stimuli were recorded beforehand using an omnidirectional microphone (Earthworks TC30 K), an A/D preamplifier with phantom power-supply (MindPrint AN/DI PRO), a digital I/O sound card (RME Digi 96/8 PST), real-time sound analysis software (wSpecGram) and presented by means of a single digital loudspeaker (Genelec S30D) placed 2 m in front of the monkey (see Fig. 1B). This equipment allows the linear reproduction of frequencies in the range of 36 Hz–48 kHz (±2.5 dB). To ascertain that the reproduced sounds had an amplitude comparable to their natural counterparts, the peak sound pressure level of natural actions was measured at the head position of the monkey using a sound level meter (Lutron SL-4001). The amplitudes of the digitally reproduced sounds were matched to the same peak sound pressure using the same sound level meter. Peak sound level varied between 60 and 85 dB. Ten slightly different versions of the sound of each action were pre-recorded. This avoids the artificiality of presenting always the same sound to the monkey from one trial to another.

Auditory response onset latency

For each neuron, response onset was defined in the sound-only condition of the best action as the time when activity exceeded spontaneous activity, i.e. >mean+1.96 SD of the 1 s before sound onset (see the dotted horizontal lines in Figs. 1, 3 in the S condition). Since in the V, V+S and M conditions, preparatory parts of the action occur prior to sound onset, true spontaneous activity could only be estimated in the S conditions, where the experimenter stood still behind the loudspeaker and nothing could tell the monkey that a particular sound was going to be played back to him.

Statistical selectivity criteria

To analyze the effect of the two modalities and the two actions on the firing rate of individual cells, a Vision (yes/no) × Sound (yes/no) × Action (best/less effective) multivariate analysis of variance (MANOVA) on the neurons' firing rate in the early epoch and the late epoch was performed. The early epoch extended 1 s before the auditory response onset to capture visual responses. The late epoch started at auditory response onset and lasted for as long as the longer of the two sounds used for testing the neuron. A neuron was considered auditory selective if it had a significant A×S or A×S×V interaction, auditory non-selective if it only had a significant main effect of sound. A neuron was called visually selective if it had a significant A×V or A×S×V interaction. Since the auditory selectivity based on the A×S and A×S×V criterion may be achieved in part due to differences in the vision-and-sound condition, Newman-Keuls post hoc analyses were performed to check differences between individual conditions. In particular, a comparison of the responses to the best and less effective action during the late epoch of the sound-only condition was used to test if a neuron could differentiate between two actions based on sound-only. For single cells, significant results refer to $p<0.05$. The same analysis was also applied to the population of neurons, with one entry per neuron rather than one per trial.

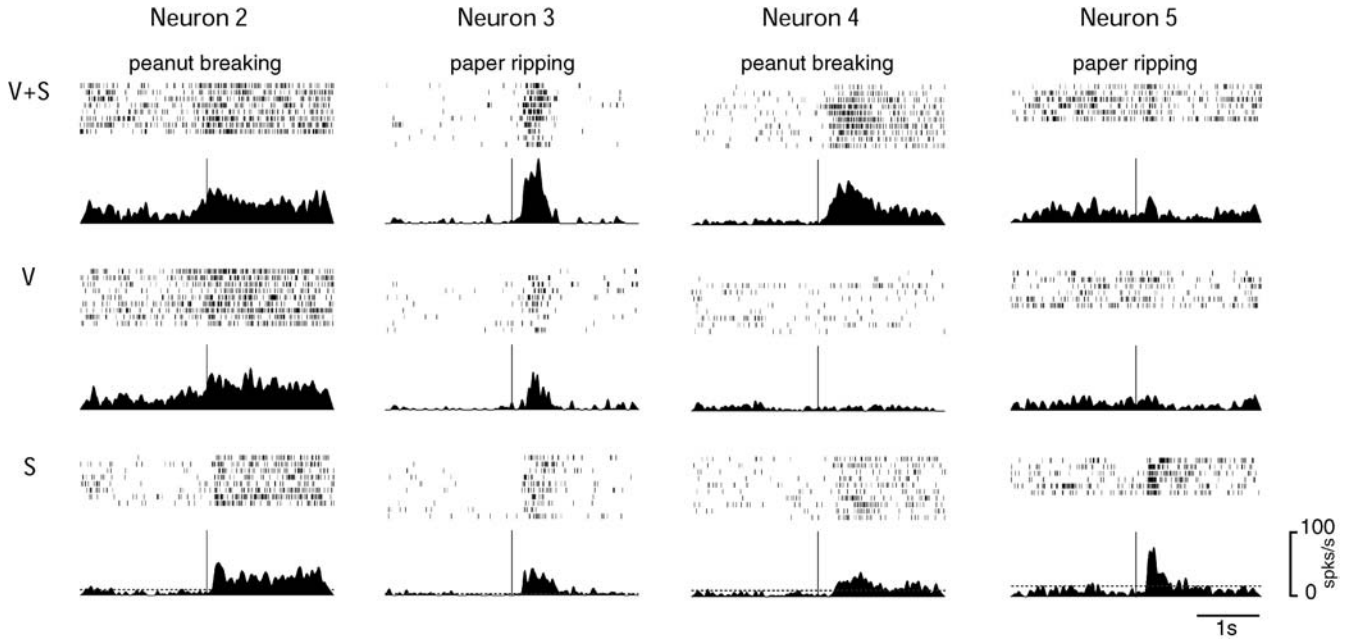

Fig. 3 Responses of four neurons to their best action stimulus in the three sensory conditions. Conventions as in Fig. 1C

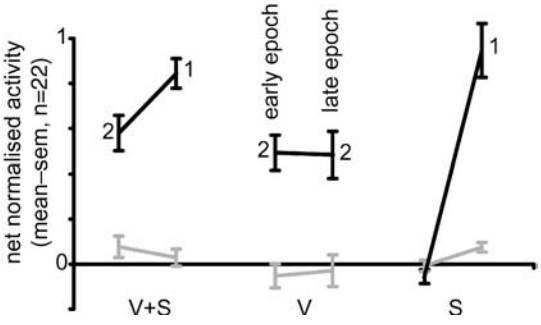

Fig. 4 Population analysis of the 22 audiovisual mirror neurons. Mean net normalized firing rates (± SEM) are shown for the early and late Epoch of the best and less effective action. *Numbers* represent the membership in one of the three homogeneous groups determined by a Newman-Keuls post hoc analysis ($p<0.05$): members of a given group do not differ significantly from each other, but do differ significantly from members of the other groups

Population analysis

Neurons differ in their response onset latency, their spontaneous activity, and their peak firing rate. To analyze the activity of the population, these factors were normalized. First, to account for differences in latency between neurons, all population analyses were done relative to the auditory response onset rather than stimulus onset. Responses were aligned on the auditory response onset even in the vision-only and motor conditions to keep all the responses aligned equally. For the first population analysis (Fig. 1D), and for each neuron, the net mean activity was calculated for each 20-ms bin, in all sensory and motor conditions and for both actions. The spontaneous firing rate (i.e., mean firing rate in the first 2 s of the sound-only conditions) was subtracted and the highest remaining bin in the best vision-and-sound condition taken to divide spike counts in all bins. In the second analysis (Fig. 4), the mean firing rate in the early and late epochs was calculated for each neuron, the spontaneous activity subtracted, and the epoch yielding the largest remaining activity in the best vision-and-sound condition taken to divide all other values for that neuron. In both analyses, zero then represents spontaneous activity, and 1 peak activity in vision-and-sound.

Receiver operator characteristic (ROC) analysis

How useful are audiovisual mirror neurons at telling us or the brain which of two actions was performed at a given moment? Could the animal use audiovisual mirror neurons to discriminate actions? Those are two questions that the ROC analysis tries to answer. For each neuron spike counts were taken in the time period extending from 1 s before auditory response onset until response onset plus the duration of the longer of the two sounds used to test that neuron. This time period contains both the purely auditory responses occurring after auditory response onset and the preceding visual responses due to the sight of preparatory parts of the action. The mean firing rate during the window of analysis was 35 spk/s (±5 SEM between neurons). For each condition (V+S, V, S and M) separately, an ROC analysis (Newsome et al. 1989 and Box 1) was performed on the spike counts to the best and the less effective actions. The same number of trials (average 8.5±0.34 trials) was used for the best and less effective action. In the sound only condition, spiking activity before auditory response onset represents spontaneous activity given that the experimenter stood still behind the loudspeaker and no sound was played back. To evaluate a chance level of ROC performance, the ROC analysis was also performed based on that activity, using spike counts taken in a time window of the same length as for the other conditions, but ending before the auditory response onset.

Results

Activity was recorded in 286 neurons. One hundred and thirty out of 286 recorded neurons responded during both motor and sensory testing. Of these, 61 appeared to have auditory properties and were selected for further testing. Thirty-three were kept long enough to perform the full testing (see "Materials and methods") for a sufficient

number of trials in all sensory and the best motor conditions. For 28 of these the monkey also performed the less effective action.

Cross-modal interactions

In Kohler et al. (2002), we showed that, of the 33 fully tested neurons, 22 showed both visual and auditory selectivity (see Neuron 1, Fig. 1C, for example). Of the remaining 11, 7 showed auditory selectivity but lacked visual selectivity, and the remaining 4 neurons responded to the sound of both actions equally. Given that the 22 audiovisual mirror neurons preferred the same action in both the visual and auditory modality, now we analyze how the two modalities interact in determining the neural responses. Although all neurons shown below also have motor responses, their motor responses will be omitted for brevity's sake.

We assessed the cross-modal interaction by performing Newman-Keuls post hoc analysis on the responses in the late epoch (ranging from the neurons' response onset in sound only conditions and lasting for the duration of the longer of the two sounds) of the most effective action for all 22 audiovisual mirror neurons. Only the late epoch was analyzed because it contained both auditory and visual responses, while the early epoch by definition only contained visual responses. Neurons were found to fall within the three categories illustrated in Fig. 3. These categories are not sharply delimited: some cells show intermediate behaviors.

The first category of neurons was characterized by the fact that responses when the vision and the sound of the action were presented together (V+S) did not differ from those to the separate presentation of the two modalities (V or S, all $p>0.05$). Half the audiovisual mirror neurons (11/22) fell into this category. Neuron 2 (Fig. 3) illustrates this behavior. For such neurons, any evidence for the action, be it auditory or visual, is sufficient to retrieve a full-blown representation of the action.

The second category of neurons was characterized by the fact that the strongest response was observed when the sound and the vision of the action were presented together. This was true for 8/22 neurons. For five of these, both sound and vision alone evoked significant responses, but the V+S response was roughly equal to the sum of the V and S response. Neuron 3 (Fig. 3) illustrates this additive behavior. The remaining three neurons showed no significant response in the V condition, but responded more in the V+S condition than in the S condition (e.g. Neuron 4, Fig. 3). For these later neurons, the conjunction of vision and sound appears critical for the response.

The third category included three neurons that responded most strongly to the S condition. Neuron 5 is an example of such a cell.

If the 22 audiovisual mirror neurons are considered as a population (see Fig. 4), a MANOVA considering the normalized net firing rate in both the early and late Epoch reveals a main effect of Action, Sound and Vision (all $R_{(2,20)}>25$, $p<10^{-5}$), and all interactions between the factors are significant (all $R_{(2,20)}>8$, $p<0.005$). A Newman-Keuls post hoc analysis revealed three homogeneous groups (at $p<0.05$). The group containing the largest values included the late epoch of V+S and S, the intermediate group contained the early epoch of V+S and the early and late epoch of V. The third group contained the remaining conditions that had activity values not differing from spontaneous activity. Values within a homogeneous group do not differ significantly from each other, while values taken from different groups do.

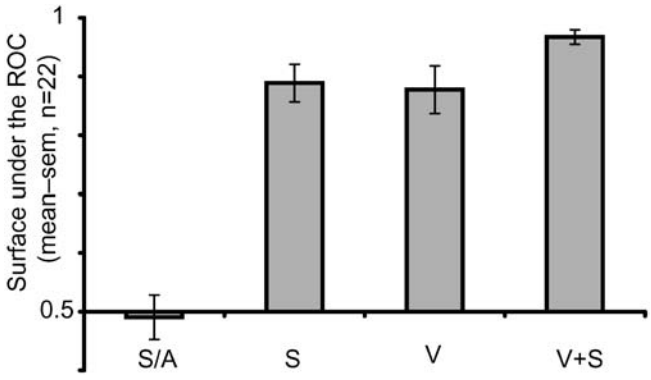

Fig. 5 Mean (± SEM) surface under the ROC curve for the 22 audiovisual mirror neurons. S/A stands for spontaneous activity and is the result of the ROC analysis if spikes are counted in the sound only condition before the sound has been played back. Other conventions as in Fig. 4

The existence of a significant Action × Sound × Vision interaction indicates that at the population level, the contributions of the visual and auditory modality were not independent. In the light of the Newman-Keuls post hoc analysis, this significant interaction is due to the fact that in the late epoch, the vision of the best action has a strong impact on the response if the sound is absent, but not when the sound is present.

Receiver Operator Characteristics (ROC) analysis

If audiovisual mirror neurons play a role in the recognition of actions, their firing rate has to reliably discriminate between actions (see Box 1; Appendix 1). To evaluate how well a monkey would perform in an action discrimination task if he used the firing of audiovisual mirror neurons as the only source of information, an ROC analysis was performed. Figure 5 shows the average surface under the ROC curve obtained for the 22 audiovisual mirror neurons. These values estimate the proportion of correct answers the monkey would be able to give if asked 'which of the two tested actions was this?', and basing its answer only on the firing of an audiovisual mirror neuron. The leftmost bar represents the average performance obtained if the analysis is based on

spontaneous activity (see "Materials and methods"), and is equal to 0.49, not differing from the expected chance level of 0.5. Performance in the V and S condition averaged at 0.88 and 0.89, respectively. In the V+S condition, performance reached 0.97. The monkey could therefore on average differentiate the two tested actions with a performance of ~90% correct based on vision or sound alone, and 97% correct based on the combined vision and sound of the action, if he/she only used the firing of a single audiovisual mirror neuron to take this decision.

A 2 vision × 2 sound repeated measurement ANOVA on these results indicates significant main effects for vision ($F_{(1,21)}$=44, $p<10^{-6}$) and sound ($F_{(1,21)}$=63, $p<10^{-6}$) as well as a significant interaction ($F_{(1,21)}$=46, $p<10^{-6}$). A Newman-Keuls post hoc analysis revealed three homogeneous groups: spontaneous activity with the smallest performance, V and S with intermediate performance, and V+S with best performance.

If only the 14 neurons were considered, for which both the best and the less effective action were tested in the motor condition, the ROC analysis yields the following results (mean ± SEM): 0.81±0.05 (M), 0.87±4 (S), 0.86±0.06 (V) and 0.97±0.02 (V+S). Despite the fact that the neurons are recorded in the premotor cortex, their firing therefore more accurately predicts what action the monkey observed (V+S) than what action the monkey performs (M, *t*-test, $p<0.01$).

Discussion

Previously (Kohler et al. 2002) we have shown that some neurons in the ventral premotor cortex (area F5) of the monkey responding during the execution of actions also respond to the vision and/or the sound of these actions. Here we show that for half of the tested audiovisual mirror neurons, the amplitude of the response does not differ significantly whether the preferred action is heard, seen or both heard and seen. We also demonstrate that the firing of audiovisual mirror neurons would support quasi-perfect sensory discrimination performance between two actions.

When we say that we *recognize* that someone just knocked on the door, we mean that we matched the sound of this action with our internal representation of what 'knocking on the door' is. A striking property of audiovisual mirror neurons is the fact that they match the sound and the vision of someone else's actions onto the monkey's own motor repertoire. It is therefore likely that these neurons participate in the recognition of an action: We recognize someone else's actions because we manage to activate our own inner action representation using mirror neurons (Gallese et al. 1996; Rizzolatti et al. 1996). In this context, it is paramount that audiovisual mirror neurons not only discriminate between actions in all modalities, but that the action producing more activity in one modality is the action producing more firing also in the other modalities. As we show in this paper, audiovisual mirror neurons have this property both if considered individually and if considered as a population.

Humans can effortlessly discriminate between the sound of someone ripping a sheet of paper and someone breaking a peanut. If audiovisual mirror neurons are to play a key role in this discrimination process, their firing rate should reliably discriminate between such actions. Here we analyzed their firing rates using the ROC analysis (Newsome et al. 1989) and show that they indeed support near-perfect discrimination performance between actions. Single audiovisual mirror neurons would enable a ~90% correct discrimination performance if the actions are either only seen or only heard. The combination of seeing and hearing the actions would lead to virtually perfect (~97% correct) performance using these neurons. While this finding is encouraging, it is true that this excellent performance is obtained for actions that have been chosen to produce particularly large and particularly small responses in individual neurons, but it is important to keep in mind that only 6 actions were used to test the neurons and only 286 cells had to be tested to find 22 cells that discriminate well between these actions. Given the considerable number of neurons in area F5, it is therefore likely—albeit not demonstrated in this paper—that for any given pair of actions, some neurons would have the tuning characteristics necessary to discriminate these actions. Altogether, these results are in agreement with the idea that audiovisual mirror neurons could play a central role in the recognition of actions. A similar mechanism has been demonstrated for the visual modality in humans (Fadiga et al. 1995; Grafton et al. 1996; Decety and Grezes 1999; Buccino et al. 2001). Inactivation studies may help us understand in the future if action discrimination performance is indeed affected if F5, the area in which audiovisual mirror neurons are found, is disrupted.

Traditionally, ROC analysis has indicated very accurate discrimination capacities between stimuli in sensory cortex (Newsome et al. 1989; Keysers et al. 2001). In premotor cortex, one might expect neurons to be correlated with the actions performed by the monkey, and not with the stimuli the monkey is perceiving. The high ROC scores we observe here were measured in monkeys not involved in any explicit motor task during stimulus presentation, and therefore suggest that premotor cortex may be involved in the representation of observed/heard actions independently of motor output. Interpreting the selective response as a preparation to interact with the perceived object of the action (e.g., the peanut) is unlikely: after the experimenter performed the actions, the monkey had no access to the objects used during the actions and was always rewarded with fruit juice. Indeed, in the 14 neurons tested also in the M condition, the firing of the neuron tells us more accurately what action the monkey observed (V+S, 97% correct) than what action the monkey performed (M, 81% correct). It should be kept in mind, however, that the actions to be tested were selected to show clearly different sensory responses and not clearly different motor responses.

The audiovisual mirror neurons reported in this paper were tested on average 8.5 times for each condition, depending on how long the neuron was kept. To evaluate the effect of small trial numbers on the ROC results, we performed a boot-strapping, considering only five trials in each conditions picked at random from the ones available. This procedure was repeated 10 times, with different trials being picked each time. The resulting performances was always below that calculated with all trials, and was on average 10% under that calculated with all trials. Using a small number of trials thus tends to *reduce* the ROC performance, and the high ROC performances obtained in this paper are therefore probably an underestimate of the ROC performance that would be obtained with an even larger number of trials.

Another remarkable property of audiovisual mirror neurons is that about half of them respond with a similar intensity of discharge whether the action is only heard, only seen or both heard and seen. This finding is important, as it suggests that the neurons code the action in an abstract way, which does not depend on the source of information (auditory or visual) from which the evidence about the presence of the action is taken. For these neurons breaking a peanut is breaking a peanut, whether the monkey saw peanut breaking or heard peanut breaking. This abstract coding in neurons situated in the ventral premotor cortex may be a precursor of the abstract properties so characteristic of human thought. Indeed, bringing together the capacity for abstract representations and auditory input, audiovisual mirror neurons may be a cornerstone in the evolution of language. The fact that they are located in F5, the area considered the monkey's precursor of Broca's area (Rizzolatti and Arbib 1998), supports this idea. Indeed the abstract action representation embodied by audiovisual mirror neurons is reminiscent of the way we use verbs in language: the verb'break' is used to represent an abstract meaning that is used in different contexts: 'I see you break a peanut', 'I hear you break a peanut', 'I break a peanut'. The verb, just as the responses in audiovisual mirror neurons, does not change depending on the context in which it is used, nor depending on the subject/agent performing the action.

How audiovisual mirror neurons acquire their remarkable properties remains to be elucidated, but it is reasonable to assume that this coupling of motor, auditory and visual properties occurs through hebbian learning (Hebb 1949; Bi and Poo 2001). Whenever the monkey breaks a peanut, two events overlap in time: neurons involved in the motor planning and execution of the movement will be active, while at the same time the monkey sees and hears the consequences of this action. The consequences will include the sound of the breaking peanut and the sight of his/her own hands performing the action. The temporal overlap of activity in the motor system and activity in the sensory areas of the brain responding to the sensory consequences of the actions are ideal conditions for hebbian associative learning. The only further requirement for such learning to occur is that a single neuron must have anatomical inputs relaying motor intentions and auditory and visual feedback. To our knowledge there is no evidence for a direct anatomical connection between area F5 and auditory cortices (M. Matelli, personal communication). The auditory information may reach F5 neurons along complex cortico-cortical routes (see Romanski et al. 1999) or even involve cortico-subcortical loops (see Fries 1984). Whatever the connection may turn out to be, once hebbian associative learning has occurred, the sound alone, the vision alone or the motor intention alone could then evoke—as observed in our experiment—firing in such neurons even if the sound or the vision originate from someone else's movements. Finally, while previous findings have shown that the ventral premotor cortex contains multimodal neurons integrating auditory and visual information (Watanabe 1992; Graziano et al. 1999), the present findings substantially extend those results by showing how multimodal integration can be used for the meaningful representation and recognition of ecologically relevant actions.

Acknowledgements The research was funded by a MIURST and an ESF Grant. E.K. was supported by a Fonds fuer Medizinische Forschung der Universitaet Zuerich fellowship, C.K. by an EU Marie-Curie Fellowship. We thank G. Rizzolatti for help and advice on all aspects of the work, G. Pavan, M. Manghi and C. Fossati for their invaluable help in computing and acoustics, S. Rozzi, M. Matelli, and G. Luppino for their anatomical advice and F. Orzi and P. Rossi for caring for the monkeys.

Appendix: Box 1: the Receiver Operator Characteristic Analysis

The fundamental issue behind the Receiver Operator Characteristic (ROC) analysis is simple: the brain contains no photoreceptors, and thus has no direct vision of the outside world—instead, it contains neurons that fire a certain number of times. The brain then has to analyze what happens in the outside world based on the firing of its neurons. The ROC analysis calculates how well an imaginary observer of the firing activity of a given neuron could be at detecting a particular target stimulus. In the case at hand, this ideal observer has to decide on each trial if a particular action (his target action or best action) was performed. He has to take this perceptual decision based on the number of spikes produced by an audiovisual mirror neuron on each individual trial.

The decision rule used by the imaginary observer of the neural activity is simple: he decides on a threshold spike-count value *theta*, and compares the count x_i on a given trial i with this threshold. If $x_i \geq theta$, the observer responds that his target action occurred. If $x_i < theta$, he reports that another action must have been performed. The decision of the observer is then compared with what action was really performed on that trial.

The working of the ROC analysis can be best explained by applying the method to two different neurons: one, which firing has nothing to do with what action the experimenter performed in front of the monkey, and one—an audiovisual mirror neuron—that fires more when its best action was performed.

Figure 2A shows the spike-count distribution for the audiovisual mirror neuron. Note how the histogram when the experimenter performed the best action (top of Fig. 2A) is shifted rightwards compared with the lower histogram when the less effective action was performed (bottom). This shift means that the neuron is more likely to produce large spike-counts when the best action is performed. In the example of Fig. 2A, the observer places his *theta* at an intermediate value (e.g., 20, dotted vertical line in Fig. 2A).

Not knowing what action was really performed by the experimenter, the observer applies his decision rule blindly to both conditions. He reports the best action when the spike count is larger than or equal to *theta* (i.e., at the right of his threshold) in both cases. In our example, he reports the best action in nine out of the ten best action trials (i.e., when the experimenter really performed the best action), and in two of the ten less effective action trials. The former 9/10 are correct decisions, called 'hits' (black bars in Fig. 2A), and his hit rate is thus 90%. The latter 2/10 are errors, called false alarms (gray bars), and his false alarm rate is thus 20%. The hit and false alarm rate depend not only on the response of the neuron, but also on the *theta* used by the observer: Placing *theta* very low (i.e., moving the dotted line leftwards), the observer would always report the best action, and both the hit and false alarm rate would approach 1. Choosing very large *theta,* he would never report the best action, and both the rates will approach 0. Figure 2B illustrates the relationship between the hit rate and the false alarm rate for this audiovisual mirror neuron as a function of *theta*. This curve, called the receiver operator characteristic curve, is obtained by testing all the possible *theta* values, plotting the hit and false alarm rates for each *theta*, and connecting all the points. For this neuron, this curve is very far away from the dotted diagonal, and the surface under the curve is close to 1 (0.96). What does this large surface mean? Given that the two histograms overlap very little, if the observer starts at a *theta*=42 (bottom left of Fig. 2B), and reduces *theta* until 23, the observer only responds with best action for best action trials, i.e., the criterion does not include trials of the lower histogram in Fig. 2A, and thus the hit rate increases without increasing the false alarm rate. Only if the observer reduces *theta* below 23 will he respond with best action also in trials where the worst action was really performed. The curve thus rises vertically until the 80% hit rate, and then moves almost horizontally towards a false alarm of 100%, covering almost all the surface in the box. This is symptomatic for cases where the neuron accurately discriminates between the two types of actions.

In contrast to the audiovisual mirror neuron, let us now consider an imaginary neuron that does not discriminate between the two actions. Figure 2C illustrates the spike counts for this neuron. Note how much overlap there is between the two histograms. Under such conditions, placing *theta* at 20 means that the observer will respond with best action 6/10 times when the best action and 5/10 times when the worst action was really performed. The result is that the hit and false alarm rate (60 and 50% respectively) are almost equal. Figure 1D illustrates the ROC curve in this case. The curve remains very close to the diagonal, meaning that a decrease of *theta* increases the hit rate, but at the cost of equally increasing the false alarm rate. The observer is essentially randomly guessing: the spike count gives him no information about what action was performed. The surface under the curve will be close to 0.5 (here 0.54), which is symptomatic for the cases when the activity of the neuron is unrelated to the action performed by the experimenter.

The surface under the ROC curve is thus an indication of the proportion of correct decisions that the observer makes, taking all the *theta*s into account. From the two examples, it is intuitive that a surface close to 0.5 represents random performance while a surface close to 1 indicates that the observer is very good at telling what action was performed. This surface then gives us valuable information about what function the neuron might have in the brain. If the imaginary observer is very good at telling what action was performed (large surface under the ROC curve), then the brain too could use this neuron to discriminate between the two actions. The neuron could thus participate in the perception of the actions. If the observer is very poor at telling the difference between the two actions using the spike count of this neuron, so would the brain be, and the neuron therefore is unlikely to be involved in the perception of the actions.

References

Bach M, Bouis D, Fischer B (1983) An accurate and linear infrared oculometer. J Neurosci Methods 9:9–14

Bi G, Poo M (2001) Synaptic modification by correlated activity: Hebb's postulate revisited. Annu Rev Neurosci 24:139–166

Buccino G, Binkofski F, Fink GR, Fadiga L, Fogassi L, Gallese V, Seitz RJ, Zilles K, Rizzolatti G, Freund HJ (2001) Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. Eur J Neurosci 13:400–404

Cantalupo C, Hopkins WD (2001) Asymmetric Broca's area in great apes. Nature 414:505

Decety J, Grezes J (1999) Neural mechanisms subserving the perception of human actions. Trends Cogn Sci 3:172–178

Fadiga L, Fogassi L, Pavesi G, Rizzolatti G (1995) Motor facilitation during action observation: a magnetic stimulation study. J Neurophysiol 73:2608–2611

Fries W (1984) Cortical projections to the superior colliculus in the macaque monkey: a retrograde study using horseradish peroxidase. J Comp Neurol 230:55–76

Gallese V, Fadiga L, Fogassi L, Rizzolatti G (1996) Action recognition in the premotor cortex. Brain 119:593–609

Grafton ST, Arbib MA, Fadiga L, Rizzolatti G (1996) Localization of grasp representations in humans by positron emission tomography. 2. Observation compared with imagination. Exp Brain Res 112:103–111

Graziano MS, Reiss LA, Gross CG (1999) A neuronal representation of the location of nearby sounds. Nature 397:428–430

Hebb DO (1949) The organization of behavior. Wiley, New York

Keysers C, Xiao DK, Foldiak P, Perrett DI (2001) The speed of sight. J Cogn Neurosci 13:90–101

Kohler E, Keysers C, Umiltà MA, Fogassi L, Gallese V, Rizzolatti G (2002) Hearing sounds, understanding actions: action representation in mirror neurons. Science 297:846–848

Newsome WT, Britten KH, Movshon JA (1989) Neuronal correlates of a perceptual decision. Nature 341:52–54

Rizzolatti G, Arbib MA (1998) Language within our grasp. Trends Neurosci 21:188–194

Rizzolatti G, Fadiga L, Gallese V, Fogassi L (1996) Premotor cortex and the recognition of motor actions. Brain Res Cogn Brain Res 3:131–141

Romanski LM, Tian B, Fritz J, Mishkin M, Goldman-Rakic PS, Rauschecker JP (1999) Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. Nat Neurosci 12:1131–1136

Umiltà MA, Kohler E, Gallese V, Fogassi L, Fadiga L, Keysers C, Rizzolatti G (2001) I know what you are doing. A neurophysiological study. Neuron 31:155–165

Watanabe M (1992) Frontal units of the monkey coding the associative significance of visual and auditory stimuli. Exp Brain Res 89:233–247