THE MIRROR SYSTEM HYPOTHESIS: FROM A MACAQUE-LIKE MIRROR SYSTEM TO IMITATION

MICHAEL A. ARBIB,[2,3,4] JAMES BONAIUTO[3,4] & EDINA ROSTA[1]

[1]*Chemistry,* [2]*Computer Science,* [3]*Neuroscience and* [4]*USC Brain Project*

University of Southern California, Los Angeles, CA 90089-2520

The Mirror System Hypothesis (MSH) of the evolution of brain mechanisms supporting language distinguishes a monkey-like mirror neuron system from a chimpanzee-like mirror system that supports simple imitation and a human-like mirror system that supports complex imitation and language. This paper briefly reviews the seven evolutionary stages posited by MSH and then focuses on the early stages which precede but are claimed to ground language. It introduces MNS2, a new model of action recognition learning by mirror neurons of the macaque brain to address data on audio-visual mirror neurons. In addition, the paper offers an explicit hypothesis on how to embed a macaque-like mirror system in a larger human-like circuit which has the capacity for imitation by both direct and indirect routes. Implications for the study of speech are briefly noted.

1. The Mirror System Hypothesis

Both premotor area F5 and parietal area PF of the macaque monkey brain contain *mirror neurons* each of which fires vigorously both when the monkey executes a certain limited set of actions and when the monkey observes some other perform a similar action. Imaging data show that the human brain contains *mirror regions* in both frontal and parietal lobes, namely regions that show high activation both when a human performs a manual action and when the human observes a manual action, but not when the human simply observes an object. It is widely assumed that such mirror regions contain mirror neurons, based on similarities between the human and macaque brain.

The *Mirror System Hypothesis* (MSH; Rizzolatti and Arbib, 1998) asserts that the *parity requirement* for language in humans – that what counts for the speaker must count approximately the same for the hearer – is met because Broca's area (often associated with speech production) evolved atop the mirror system for grasping with its capacity to generate and recognize a set of actions. However (Hurford, 2004), one must distinguish the mirror system for the signifier (phonological form) from the neural schema for the signified, and note the need for linkage of the two. On this view, Broca's area becomes the meeting place for phonological perception and production, but other areas are required to link phonological form to semantic form.

The crucial point is that humans have capacities denied to monkeys. Mirror regions in a human can be activated when the subject imitates an action, or even just imagines it, but there is a consensus that monkeys cannot imitate save in the most rudimentary sense. By contrast, chimpanzees exhibit "simple imitation", the ability to approximate an action after observing and attempting its repetition many times; while humans alone among the primates have the capacity for "complex imitation", being able to recognize another's performance as a combination of more-or-less familiar actions and to use this recognition to approximate the action, with increasing practice yielding increasing skill. Thus research on MSH requires not only a fuller understanding of the mirror system of the macaque, but also an understanding of how the mirror system and the circuitry with which it interacts must have changed in the course of evolution.

Arbib (2002, 2005a) modified and developed MSH by hypothesizing seven stages in the evolution of language. The first three stages are pre-hominid:

S1: Grasping.

S2: A mirror system for grasping, shared with the common ancestor of human and monkey.

S3: A system for simple imitation of grasping shared with the common ancestor of human and chimpanzee.

The next 3 stages distinguish the hominid line from that of the great apes:

S4: A complex imitation system for grasping.

S5: *Protosign*, a manual-based communication system that involves the breakthrough from employing manual actions for praxis to using them for pantomime (not just of manual actions), and then going beyond pantomime to add conventionalized gestures that can disambiguate pantomimes.

S6: *Protospeech*, resulting from linking the mechanisms for mediating the semantics of protosign to a vocal apparatus of increasing flexibility.

Arbib (2005b) argues that protosign and protospeech evolved together in an expanding spiral. The final stage is then:

S7: *Language*: the change from action-object frames to verb-argument structures to syntax and semantics.

Arbib (2005) provides arguments and counter-arguments for these various claims. The present article focuses on the earlier, rather than the later, stages in this progression. It contributes to this argument by (a) introducing a new model of action recognition learning by macaque mirror neurons which addresses data on auditory input; (b) outlining how to embed a macaque-like mirror system in

a larger human-like circuit which has direct and indirect paths for "complex imitation"; and (c) noting implications for the study of speech.

2. MNS2: Recognizing Audible Actions

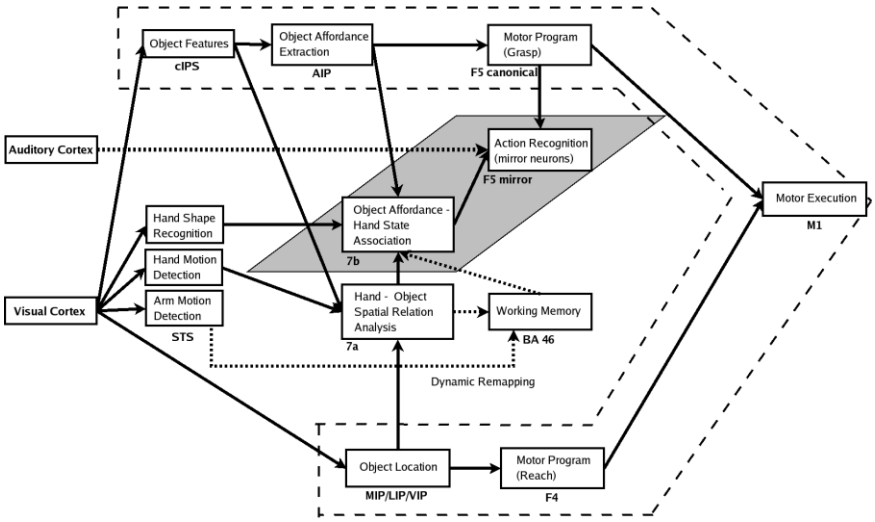

Figure 1: System diagram for the MNS2 model, updating the MNS model of Oztop & Arbib, 2002. The dashed outline shows the system for generating the reach to and grasp of an observed object. The remaining circuitry defines the mirror system and the subsystems which feed it. The encoding of the grasp motor program (F5 canonical) provides the training signal for a recurrent network which models the areas 7b and F5 mirror, shown here in the gray parallelogram, by the activity of its hidden and external output layers, respectively. The dotted arrows denote the connections unique to the MNS2 model. Auditory information about actions reaches the F5 mirror neurons via the auditory cortex. Visual data on hand-object spatial relations is input into the Object Affordance-Hand State Association schema and into working memory. When this information is not available externally, the dynamically remapped working memory trace serves in its place.

The MNS model (Oztop & Arbib, 2002) of the monkey mirror system was designed to associate activity in canonical neurons providing a premotor encoding of the type of a grasp with visual input encoding the trajectory of a hand relative to an observed object. The learning mechanism was a feed-forward backpropagation network of units with one hidden layer which required an unnatural recoding of its input. Bonaiuto et al. (2005) developed a model, MNS2, that could process the time series of hand-object relationships without such recoding, using an adaptive recurrent network to learn to classify grasps based on the temporal sequence of hand-object relations.

Umiltá et al. (2001) have shown that mirror neurons in the macaque monkey can recognize a grasp if the monkey has seen the target object which

was then hidden, but cannot recognize the action lacking current or recent input on the affordances and location of the object. MNS2 incorporates working memory and dynamic remapping components (Figure 1) which allow the model to recognize grasps even when the final stage of object contact is hidden and must be inferred. Before being hidden, the object position and its affordance information are stored in working memory. Once the hand is no longer visible, the working memory of wrist position is updated using the still-visible forearm position. If the model observes an object which is then hidden by a screen, and then observes a grasp that disappears behind that screen, the wrist trajectory will be extrapolated and the grasp will be classified accordingly.

However, the more important contribution of MNS2 within the context of MSH is that it addresses data on "audiovisual" mirror neurons which associate sounds with manual actions. Köhler et al. (2002 –see Figure 2 right) found that some of the mirror neurons in area F5 of the macaque premotor cortex responsive to the sight of actions associated with characteristic noises (such as peanut breaking) are just as responsive for the sounds of these actions.

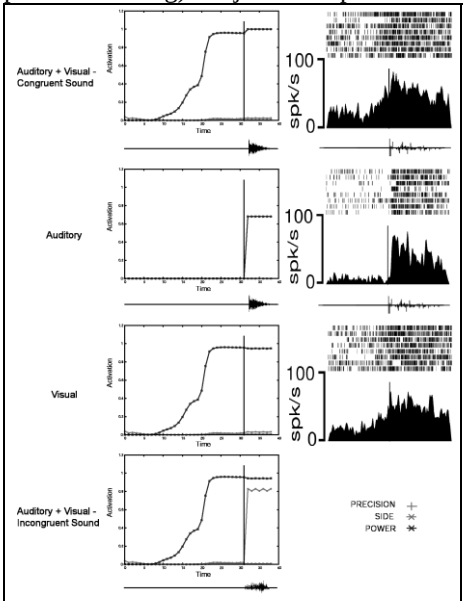

Figure 2: Left: Activation of the model's external output layer when presented with a precision grasp sequence containing (from top to bottom) visual and congruent audio, visual only, audio only, and visual and incongruent audio information. The black vertical lines indicate the time step at which the hand made contact with the object. The unit encoding the precision grasp shows the greatest level of activation, while the unit corresponding to power grasps shows a small level of transient activity. At the bottom is an oscillogram of the sound associated with the precision grasp. Right: Activation from Köhler et al. (2002) of an audiovisual mirror neuron responding to (from top to bottom) the visual and audio components, visual component alone, and audio component alone of a peanut-breaking action.

Bonaiuto et al. (2005) associate each sound with a distinct pattern of activity which is applied to audio input units which are fully connected to the output layer of the recurrent neural network, corresponding to a direct connection from auditory cortex to F5. These connection weights are modified using Hebbian learning. In this way, any sound that is consistently perceived

during multiple occurrences of an executed action becomes associated with that action and incorporated into its representation. This type of audio information is inherently actor-invariant and this allows the monkey to recognize that another individual is performing that action when the associated sound is heard.

3. A Dual Route Model of Imitation Gated by Attention

It is often suggested that mirror neurons are the substrate for imitation, matching observed actions onto motor programs producing similar or equivalent actions. However, as we saw earlier, only humans have "complex imitation", the ability to imitate sequences of behaviors and approximate novel actions as variants of known actions after one or just a few viewings of this novel behavior. As backdrop for our own work, we draw some important lessons from apraxia.

DeRenzi (1989) reports that some apraxics exhibit a *semantic deficit* – having difficulty both in classifying gestures and in performing familiar gestures on command – yet may be able copy the pattern of a movement of such a gesture without "getting the meaning" of the action of which it is part. We call this residual ability *low-level imitation* to distinguish it from imitation based on recognition and "replay" of a goal-directed action. With Rothi, Ochipa, and Heilman (1991), we thus propose a dual route imitation learning model to serve as a platform for studying apraxia. The *direct route* for imitation of meaningless and intransitive gestures converts a visual representation of limb motion into a set of intermediate limb postures or motions for subsequent execution. The *indirect route* for imitation of known transitive gestures recognizes and then reconstructs known object-directed actions. The distinction between the direct and indirect routes in praxis may be related to the well-known distinction between the dorsal and ventral streams in vision (Ungerleider & Mishkin, 1982) which also plays a crucial role in our model of the visual control of hand movements (Fagg & Arbib, 1998) and may in turn have implications for the study of language. We suggest that the interaction of these two routes underlies the human capacity for complex imitation. We hypothesize that, during sequential or complex actions, contributions from each route are encoded in a competitive queuing mechanism (Rhodes et al., 2004). The focus of attention (whether directed toward the object and limb, limb posture, or movement) determines the relative competitive weight of the movement segment encoded by each route. A modification to the competitive choice layer implements a sort of selective, n-winners-take-all mechanism,

allowing non-interfering movement segments with similar weights to be executed simultaneously. In this way novel movements can be recognized as combining known actions (indirect route) with intransitive limb adjustments (direct route).

4. Complex/Goal-Directed Imitation

We have argued that humans have "complex imitation", the capacity for recognizing novel actions as combinations of (variants of) known object-directed actions, with joint adjustments to meld them together. These novel actions can then be acquired as skills through successive approximation. In addition, humans have the ability to imitate complex "meaningless" movements which are not directed towards objects – as we saw in defining the "direct route".

In their theory of goal-directed imitation, Wohlschläger et al. (2003) present the hypothesis that imitation is the result of the decomposition of the aspects of a movement and the hierarchical structuring of these goal aspects. Each of these goal aspects triggers the associated motor program for reproducing that aspect of the movement. Wohlschläger et al. (2003) attribute differences in imitative abilities across species to differences in working memory capacity. However, this is not evident from the current data, and differences in imitative ability could very well be due to differences in the mechanism(s) of hierarchical movement aspect decomposition. The fact that humans can imitate intransitive movements does not seem to be due to an increased working memory capacity, but rather the ability to decompose aspects of intransitive movements such as relative limb postures and via points. Through this process of successive approximation, complex movements can be reproduced with increasing accuracy by increased attention being paid to its subparts. This increased attention may result in a finer-scaled decomposition of the observed movement, resulting in execution of a more congruent movement.

5. Discussion

The audio properties of mirror neurons are of major interest because they may have been crucial in the transition from gesture to vocal articulation in the evolution of language. These multi-modal mirror neurons may have allowed arbitrary vocalizations to become associated with communicative gestures, facilitating the emergence of a speech-based language from a system of manual gestures. If this is indeed the case, the development of audio-visual mirror neurons may have implications for the recognition of communicative actions

and ground the multi-modality of language (Fogassi & Ferrari, 2004; Arbib, 2005b).

The possible relation of the direct and indirect routes in praxis to the dorsal and ventral streams in vision may in turn have implications for the study of language. Hickok & Poeppel (2004) observe that early cortical stages of speech perception involve auditory fields in the superior temporal gyrus bilaterally (although asymmetrically) but offer evidence that this cortical processing system then diverges into two streams:

A *dorsal stream* maps sound onto articulatory-based representations which projects dorso-posteriorly. It involves a region in the posterior Sylvian fissure at the parietal–temporal boundary, and ultimately projects to frontal regions. This network provides a mechanism for the development and maintenance of "parity" between auditory and motor representations of speech; and

A *ventral stream* maps sound onto meaning which projects ventro-laterally toward inferior posterior temporal cortex (posterior middle temporal gyrus) which serves as an interface between sound-based representations of speech in the superior temporal gyrus (again bilaterally) and widely distributed conceptual representations.

The distinction between the direct and indirect routes in praxis may also be relevant to the distinction made by Levelt (e.g., Levelt et al., 1999) between overt and internal speech. Using our normal perceptual system, we can monitor our own vocal output and discover errors, dysfluencies, or other problems of delivery in our own overt speech. However, Levelt further claims that we can monitor some internal representation – Wheeldon and Levelt (1995) offer evidence that this takes the form of a somewhat abstract phonological representation – as it is produced during speech encoding and use this internal self-monitoring ability to trace the process of phonological encoding itself. As noted by one of the reviewers, a fruitful topic for future research is to pursue the development of this dual-feedback architecture on an evolutionary scale as part of the task of elaborating the Mirror System Hypothesis.

References

Arbib, M.A. (2005a). From Monkey-like Action Recognition to Human Language: An Evolutionary Framework for Neurolinguistics, *Behavioral and Brain Sciences, 28*, 105-167. (Supplemental commentaries and the author's "electronic response" are at *Behavioral and Brain Sciences,* http://www.bbsonline.org/Preprints/Arbib-05012002/Supplemental/Arbib.E-Response_Supplemental.pdf.)

Arbib, M.A. (2005b). Interweaving Protosign and Protospeech: Further Developments Beyond the Mirror, *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems 6*, 145–171.

Bonaiuto, B., Rosta, E., and Arbib, M.A. (2005). Recognizing Invisible Actions, *Workshop on Modeling Natural Action Selection*, Edinburgh, July, 2005. (An expanded version has been submitted for publication under the title "Extending the Mirror Neuron System Model, I: Audible Actions and Invisible Grasps".)

DeRenzi, E. (1989). Apraxia. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology*. Amsterdam: Elsevier. Vol. 2., pp. 245–263.

Fagg, A.H., Arbib, M.A. (1998). Modeling Parietal-Premotor Interactions in Primate Control of Grasping, *Neural Networks 11*, 1277-1303.

Fogassi, L., Ferrari, P.F. (2004). Mirror neurons, gestures and language evolution, Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems 5,345-363.

Hickok, G., and Poeppel, D., 2004, Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language, *Cognition*, *92*, 67-99.

Hurford, J.R. (2004). Language beyond our grasp: what mirror neurons can, and cannot, do for language evolution. In D.K. Oller and U. Griebel (Eds.), *Evolution of Communication Systems: A Comparative Approach*. Cambridge, MA: The MIT Press, pp. 297-313.

Köhler, 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.

Levelt, W.J.M., Roelofs, A., Meyer, A.S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1-38.

Oztop, E., Arbib, M.A. (2002). Schema design and implementation of the grasp-related mirror neuron system. Biological Cybernetics, *87*, 116-140.

Rizzolatti, G., Arbib, M.A. (1998). Language Within Our Grasp. *Trends in Neuroscience*, *21*, 188-194.

Rothi, L.J.G., Ochipa, C., and Heilman, K.M. (1991). A cognitive neuropsychological model of limb praxis. *Cogn. Neuropsychol*. *8*, 443-458.

Umiltá, M.A., Köhler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., Rizzolatti, G. (2001). I know what you are doing: a neurophysiological study. *Neuron*, *31*, 155-65.

Ungerleider, L.G., Mishkin, M. (1982) Two cortical visual systems. In D.J. Ingle, M.A. Goodale, and R.J.W. Mansfield (Eds.), *Analysis of Visual Behavior*. Cambridge, MA: The MIT Press, pp.549-586.

Wheeldon, L.R., Levelt, W.J.M. (1995) Monitoring the time course of phonological encoding. *Journal of Memory and Language 34*, 311–34.

Wohlschläger, A., Gattis, M., Bekkering, H. (2003). Action generation and action perception in imitation: an instance of the ideomotor principle. *Phil. Trans. R. Soc. Lond., 358*, 501-515.