Imitation in embodied communication – from monkey mirror neurons to artificial humans

Stefan Kopp and Ipke Wachsmuth
Artificial Intelligence Group, Faculty of Technology
University of Bielefeld
{skopp,ipke}@techfak.uni-bielefeld.de

James Bonaiuto[1,3] and Michael Arbib[1,2,3]
[1]Neuroscience, [2]Computer Science, and [3]USC Brain Project
University of Southern California
bonaiuto@usc.edu, arbib@pollux.usc.edu

1. Introduction

The notion of *Embodied Communication* put forward in this book emphasizes the role of the body and the perceptuo-motor system in communication and social cognition. One obviously important aspect of human-human communication is the integral use of the body in orchestrating words and prosody with hand gesture, facial expressions or body posture for multimodal information transfer. Another aspect is the possible role of lower cognitive levels of perceptual and motor processing in establishing what may be called "empathic couplings" between interacting individuals. Such couplings are revealed, e.g., by the "chameleon effect" (Chartrand and Bargh 1999), the tendency of humans to non-consciously mimic the postures and movements observed in others.

Beyond mimicry, the term "imitation" has been used to denote a variety of phenomena. We start by taking it to include all cases where an individual performs a behavior that resembles a behavior previously performed by others in a communicative interaction. We will elucidate the different kinds of imitative behavior below, but note here that the empathic couplings we see in embodied communication exemplify a link between communication and, in this case, non-conscious imitative behavior. Human communication, however, also rests on a shared set of symbols – words and morphemes – and the constructions that specify the combinations of form and meaning, whereby words are combined into phrases and phrases into sentences. Each of us acquires this stock in trade of language both through conscious imitation of the sounds of words and the use of phrases and through further subconscious assimilation of diverse shadings and usage. In short, whether consciously or non-consciously, the process of *imitation* is critical whether in acquiring the symbolic structure of a language or in the everyday interactions of communication that involve so much more than exchanging strings of symbols.

When imitating someone's actions we perceive a behavior in a structured way and transcode it into motor commands from our own repertoire. This process requires, and is the main mechanism by which we acquire, what we suggest to be at the basis of everyday (embodied) communication: an automatic connection of the perception of others with our own motor and personal knowledge. The idea is that the observation of others' behavior may serve both to prime specific representational structures involved in the generation of our own communicative behavior, and to generate inter-personal couplings that coordinate and align the interlocutors below the level of their intentional contributions. Although such non-conscious activations may generally also lead to "decouplings" – an antagonistic bias that may impede successful communication when one speaker gets a bad impression of another – we will here focus on "sympathetic" inter-personal couplings.

Different accounts of imitative behavior have proposed that the ability to connect a perceived movement with the own action repertoire rests upon a, more-or-less aware, attention to different *significances* of the observed behavior, from the familiarity of motor level acts to the effectivities of skilled action for the achievement of task-level goals. Those recognized significances are assumed to be associated with our own experiences and intentions, which could then be either simply passed to our own motor action system for imitation or behavioral priming, or could be consciously exploited to infer the prior or social intentions of others. Many researchers thus have come to believe that we understand others by internally simulating them, employing mechanisms of imitation in a covert fashion (Gallese and Goldman 1998; Wilson and Knoblich 2005); but see (Jacob and Jeannerod 2005) for a critical discussion.

This, however, is only part of the story. When we engage in communication with another we take turns and, when acting as the hearer, we respond to the speaker in two different ways: the first is to update the shared context for the conversation, and here some form of covert imitation of the speaker may serve to interpret and assimilate their message into this evolving context; the second is to determine (not necessarily consciously) how this shared context coupled with one's communicative goals will build on the speaker's current "move" to yield one's own utterance when taking the turn as speaker. In some cases, repetition of part of what has just been heard or observed may be appropriate, e.g., to confirm an instruction. This is also a normal occurrence in childhood development. Note, however, in older children and adults the immediate and involuntary repetition of words or phrases just spoken by others is a disorder, echolalia, and may occur as a symptom of autism or some types of schizophrenia (Williams *et al.* 2001).

In this paper, we approach the issue of which roles imitation plays in and for embodied communication from two different directions. In the first, we look at the "mirror system" of the macaque brain, assessing models of neurons, which are active both when the monkey itself acts in a certain way, and when the monkey sees another monkey or a human executing a similar action. Monkeys have little in the way of skills in imitation, but we will trace an evolutionary path that leads from mirror neurons via increasingly complex forms of imitation to forms of embodied communication including language. In the second approach, we start with a "virtual human" – whose "body" exists only as a simulation on a computer screen – seeking to make computationally explicit the ways in which enabling such artificial agents to engage in and learn through imitation can help them attain to better capabilities of communicating with humans, by combining speech output with appropriate gestures of hands, face and body.

In both efforts we attempt to tackle an account of the role of imitation, its underlying functions and mechanisms, in *communicative* behavior. Many studies have been carried out to explore mechanisms for imitation of "transitive", i.e., object-directed, actions like grasping a cup or cracking a nut. Such movements can be perceived and conceptualized in terms of object manipulations or hand-object relations. Only recently has research begun to investigate to what extent this account can be extended to bodily movements that are primarily communicative.

2. Mechanisms of imitation

The term "imitation" has been used to denote phenomena where an individual under certain situational circumstances performs an action that resembles those previously performed by others. This can include a large variety of behavior, in all of which the brain needs to solve the so-called "correspondence problem" (Brass and Heyes 2006), mapping from some visual stimuli of another's action onto motor representations of its body sufficient to produce a corresponding action. We consider imitation to rest on some measure of analysis to the form of movement, and we distinguish it from *emulation* where the observer attains the same goals as the observed actor, but not using the same specific acts. Also, we distinguish it from *stimulus enhancement* where the mere recognition of an object primes the observer to find a way to act on that object. On the other hand, we speak of *effector enhancement* when observation of an effector used (e.g., hand versus mouth) primes the choice of effector for that found action. *Response facilitation* is the seemingly automatic, selective enhancement of a motor response when observing someone performing a corresponding act. It is often suggested as basis of a capacity for immediate imitation, and it is likely to underlie the effects of co-activation and non-conscious mimicry we take as indicative of empathic couplings as mentioned above.

It can be debated whether these are diverse forms of imitation or just imitation-like, and according to (Tomasello *et al.* 1993), *true imitation* is only present when the imitated behavior is novel for the imitator and learned by precisely reproducing the task strategy to accomplish the same goal. This, however, conflates two different issues – the level of observation of the other required in carrying out a subsequent performance (based perhaps on working memory of certain parameters) and the ability to add a new action to one's repertoire on the basis of one or more performances of this type. Moreover, the characterization of a task strategy is a subtle one, since it may rest on the way in which actions are related to the achievement of subgoals. For example, one might imitate a given action on two different occasions by "twisting cap three times" and then "twisting cap four times", but only have truly added this action to one's repertoire when one can execute each of these alternatives as an instance of "twist the cap until the cap separates from the bottle". We thus distinguish *acting by true imitation* based on the attempted reproduction of the subgoals and movements of a recently observed performance, from *learning by true imitation,* which adds a new action (matching of goals, subgoals and movements) to the repertoire through careful observation of one or more performances of the action by others (Thorpe 1956). Thus, *acting* by true imitation will include cases where the attempted reproduction is less than perfect. It may take one trial, a few trials, or many trials for such action to yield mastery of the structure of the action, and even more trials before execution of the action without observation of a model becomes truly skilled.

Interestingly, even new-born infants can perform certain acts of imitation (e.g., poking out the tongue when an adult pokes out his tongue in front of the infant (Meltzoff and Moore 1977)) but this capacity for *neonatal imitation* is qualitatively different from that for true imitation and is, we suggest, more akin to effector enhancement.

2.1 Action recognition through the body – the mirror system

The system of the macaque brain for visuomotor control of grasping has its premotor outpost in an area called F5, which contains a set of neurons, *mirror neurons*, such that each one is active not only when the monkey executes a specific grasp but also when the monkey observes a human or other monkey execute a more-or-less similar grasp (Rizzolatti *et al.* 1996). Some preliminary data on neurosurgery patients suggest that the human brain, too, contains mirror neurons (Marco Iacoboni, personal communication to last author, July, 2007). Most relevant data on the human brain have come from imaging and transcranial magnetic stimulation (TMS) which demonstrate the presence of *mirror systems*, parts of the motor system which are active when performing a certain class of actions and when perceiving such actions performed by other individuals; see (Rizzolatti and Craighero 2004) for a review.

The response of a mirror neuron when a *monkey* observes another's hand action always requires a "transitive" action, i.e., an interaction between a biological effector and an object (or, at least, very recent viewing of the now occluded object, (Umiltá *et al.* 2001)). (See, however, (Ferrari *et al.* 2003) for data on mirror neurons responding to the observation of ingestive and communicative mouth actions.) The monkey's viewing of an intransitive hand action or the intransitive mimicking of a transitive hand action are ineffective with respect to their excitation. Moreover, the firing of some mirror neurons may correlate with the goals or consequences of an action (e.g., for grasping an object with jaws or either hand). This, however, is far from a full specification of the action. During self-action, the overall motor system must specify all the details of execution of the current action. This goes beyond what can be coded by a single neuron, and more data are needed on the total information about an observed action that can be encoded by the firing of a whole population of mirror neurons.

Audiovisual neurons are responsive to the sound as well as the sight of actions (Kohler *et al.* 2002) so long as that sound is distinctive as in breaking a peanut or tearing paper. Clearly, in this case the mirror neurons can characterize only the type of action, not the particular movements needed to execute it – yet the latter are necessary for imitation of the particular movement. The monkey mirror system is thus believed to be at the basis of action recognition, but as far as hand actions are concerned, is in general restricted to transitive actions that are already in the repertoire of the observing individual.

Notwithstanding, just as the firing of mirror neurons can become associated with the sounds of actions, when these are distinctive, so can other kinds of extensive experience broaden the range of conditions that lead a mirror neuron to fire. For example, (Ferrari *et al.* 2005) found that if monkeys had extensive experience in watching a human experimenter grasp an object with a tool such as a pair of pliers, then some of the mirror neurons initially responsive to grasping with the hand would also become responsive to the grasping with the pliers. We would suggest that this involved generalizing the significancies of the relevant end-effector to include the jaws of the pliers as well as the opposed thumb and fingers of the hand.

Brain imaging data have shown that human mirror system activation can be evoked by *intransitive* as well as transitive movements (Rizzolatti and Craighero 2004), and the course of temporal excitability during action observation suggests that the *movements* forming an action are coded too. Moreover, where monkeys have little ability for imitation, humans of course do – and mirror system activity occurs during imitation (and even imagination) as well as action recognition. Psychological experiments showed that observing a given finger movement facilitates the execution of a similar movement; the greater the similarity the stronger is the priming (Brass *et al.* 2000; Craighero *et al.* 2002). These findings suggest that perception and action have in common a representation that can be used for planning and controlling the imitator's response (Knoblich and Prinz 2005). In communication, however, we generally respond to one action with a different one. Thus, the brain activity when "observing" or "preparing to imitate" does not exhaust the brain activity when "preparing to respond" in communication, a distinction that must be kept in mind. (See the chapter by Sebanz and Knoblich, this volume, for a related discussion.)

2.2 The mirror system hypothesis

We have seen that area F5 in the monkey is the premotor area containing mirror neurons. The homologous region of the human brain is Brodmann's Area 44, part of Broca's area, which is traditionally thought of as a speech area. However, this area has been shown by brain imaging studies to be also active when humans both execute and observe grasps. It is posited that the mirror system for grasping was also present in the common ancestor of humans and monkeys (perhaps 20 million years ago) and that of humans and chimpanzees (perhaps 5 million years ago). Moreover, the mirror neuron property resonates with the *parity requirement* for language – that what counts for the speaker must count approximately the same for the hearer. In addition, normal face-to-face speech involves manual and facial as well as vocal gestures, while signed languages are fully developed human languages that do not involve vocalization. These findings ground:

The Mirror System Hypothesis (Arbib and Rizzolatti 1997; Rizzolatti and Arbib 1998): *The **parity requirement** for language in humans is met because Broca's area evolved atop the mirror system for grasping which provides the capacity to generate and recognize a set of actions.*

Recent work (see (Arbib 2005a) for a review, and commentaries on current controversies) has elaborated the Hypothesis, defining an evolutionary progression of seven stages, S1 through S7:

S1: Cortical control of hand movements.

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

A mirror system does not provide imitation in itself. A monkey with an action in its repertoire may have mirror neurons active both when executing and observing that action. However, the monkey does not repeat the observed action nor, crucially, does it use observation of a novel action to add that action to its repertoire. Thus, evolution embeds a monkey-like mirror system in more powerful systems in the next two stages.

S3: A *simple* imitation system for grasping, shared with common ancestor of human and apes.

S4: A *complex* imitation system for grasping.

Both simple and complex imitation are true imitation in the sense outlined above, but we need to clarify the distinction envisioned here. *Complex imitation* has two parts: (i) the ability to perceive that a novel action may be approximated by a composite of known actions associated with appropriate subgoals; and (ii) the ability to employ this perception to perform an approximation to the observed action, which may then be refined through practice. Both parts come into play when the child is learning a language; the former predominates in adult use of language as the emphasis shifts from mastering novel words and constructions to finding the appropriate way to continue a dialogue. We contrast this with *simple imitation* as exemplified by the finding that chimpanzees took 12 or so trials to learn to “imitate” a behavior in a laboratory setting, focusing on bringing an object into relationship with another object or the body, rather than the actual movements involved (Myowa-Yamakoshi and Matsuzawa 1999). Turning to another species of great apes, (Byrne and Byrne 1993) found that gorillas learn complex feeding strategies but may take months to do so. Teaching is virtually never observed in apes (Caro and Hauser 1992) and the young seem to look at the food, not at the methods of acquisition (Corp and Byrne 2002). Moreover, chimpanzee mothers seldom if ever correct and instruct their young (Tomasello 1999). The challenge for acquiring such skills is compounded because the sequence of “atomic actions” – e.g., the various grasps that the chimpanzee must execute to successfully manipulate the food and prepare it for eating – varies greatly from trial to trial. (Byrne 2003) implicates *imitation by behavior parsing*, a protracted form of statistical learning whereby certain *subgoals* (e.g., nettles folded over the thumb) become evident from repeated observation as being common to most performances. Apparently, the young ape, over many months, may acquire the skill by coming to recognize the relevant subgoals and derive action strategies for achieving subgoals by trial and error.

However, the ability to learn the overall structure of a specific feeding behavior over many observations is very different from the human ability for complex imitation. We will say more about complex imitation in Section 2.4. Here we analyze how imitation for praxic action relates to imitation for intentional communication in the further development of the Mirror System Hypothesis (Arbib 2005a).

We have seen that the monkey and the human brain share the capability to recognize a goal-directed action from visual stimuli, but that the human brain can recognize acts which are not necessarily tied to a transitive goal of manipulating something. It is worth noting that apes do have an ability for gestural communication and so must have some ability for recognition of intransitive, communicative actions. (Tomasello and Call 1997) hypothesize that chimps develop group-specific gestures through a process of social learning called *ontogenetic ritualization*. During this process, individuals create a communicatory signal that is not transitive, by shaping each other's behavior in repeated reciprocal interactions[1].

It may seem counter-intuitive that a more advanced imitation system is required to support imitation of seemingly more primitive actions, which involve simply a movement without an explicit goal object. However, it may be that transitive movements directed toward objects are simpler to encode computationally because the coordinate frame may be fixed on the object. Intransitive movements often involve spatial configurations of limbs whose relative positions must be within certain constraints. This involves computations in multiple, moving coordinate frames and is more computationally intensive. Thus distinct machinery is required to perform imitation of arbitrary intransitive movements (we will discuss some of these mechanisms in Sect. 3.2 and 4), and this ability presumably developed at a later stage in primate evolution than imitation and recognition of transitive movements.

The fact that monkey vocalizations are innately specified (though occasions for using a call may change with experience) – whereas a group of apes may communicate with novel gestures, perhaps acquired by ontogenetic ritualization – supports the hypothesis that it was gesture, rather than vocalization (Seyfarth *et al.* 2005) that created the opening for communication to be greatly expanded once complex imitation had evolved for practical manual skills:

S5: *Protosign*, a manual-based communication system breaking through the fixed repertoire of primate vocalizations to yield an open repertoire.

The transition from complex imitation and the small repertoires of ape gestures (perhaps 10 or so novel gestures shared by a group) to protosign involves *pantomime,* first of grasping and manual praxic actions then of non-manual actions (e.g., flapping the arms to mime the wings of a flying bird), and *conventional gestures* that simplify, disambiguate (e.g., to distinguish "bird" from "flying") or extend pantomime.

Pantomime transcends the slow accretion of manual gestures by ontogenetic ritualization, providing an "open semantics" for a large set of novel meanings (Stokoe 2001). However, such pantomime is inefficient – both in the time taken to produce it, and in the likelihood of misunderstanding. Conventionalized signs extend and exploit more efficiently the semantic richness opened up by pantomime. Processes like ontogenetic ritualization can convert elaborate pantomimes into a conventionalized "shorthand", just as they do for praxic actions. In any case, protosign comprises a system of conventionalized signs – pantomime, it is claimed, supported the emergence of protosign, but is not itself part of it.

This capability for protosign – rather than elaborations intrinsic to the core vocalization systems – may then have provided the essential scaffolding for protospeech and evolution of the human language-ready brain (Arbib 2005b). Interestingly, there are cases where for bonobos (but not chimpanzees), combining gestures with facial/vocal signals, added to the behavioral impact of the communicative act on the recipient (Pollick and de Waal 2007).

S6. *Protolanguage* as Protosign and Protospeech: an expanding spiral of conventionalized manual, facial and vocal communicative gestures.

With this, a brain that supports the multimodal production and understanding of language was established. This provides the basis for:

S7: *Language*: the development of syntax and compositional semantics.

[1] In short, in ontogenetic ritualization, an individual A performs a behavior X to *physically* elicit B's reaction Y. Eventually, B responds with Y as soon as he observes some initial portion X' of X. In due course, A produces a ritualized form of X', rather than all of X, in order to *communicatively* elicit Y.

The final stage – the transition from protolanguage to language – may have involved further biological evolution, but may instead result from cultural evolution (historical change) alone (Arbib 2005a). The question of the transition to language remains hotly debated, see e.g. (Pinker and Bloom 1990). Here, we note the importance of complex imitation for language, even though we hypothesize that the capacity for it initially evolved within the context of manual praxis.

2.3 Modeling the mirror system

We now describe our work in modeling the neural mechanisms of the first few stages of the mirror system hypothesis. This will serve both to motivate future bottom-up, neural modeling of simple and complex imitation as well as to provide a contrast with efforts directed towards a more top-down modeling of imitation in order to endow a virtual human with embodied forms of communication. These models are

FARS – a model of primate grasping
MNS – a model of the monkey mirror system, and
ACQ – a model of action selection.

These models correspond to stages S1-S2 of the mirror system hypothesis and lay the groundwork for current work in extending these models to simple imitation (S3).

2.3.1 Manual action control

The FARS model (Fagg and Arbib 1998) of primate grasping addresses the selection and execution of an appropriate grasp. It is organized around a path from a parietal region called AIP (anterior intraparietal sulcus) to a set of F5 premotor neurons, located adjacent to F5 mirror neurons and known as *canonical neurons*, to the primary motor cortex M1 that helps to control the muscles of the hand and modulate the movement of the arm:

AIP → F5canonical → M1

This path by itself mediates the choice of grasps based purely on recognition by AIP of the visible *affordances* (Gibson 1979) of objects, i.e., visual cues concerning what parts of the objects are graspable. Crucially, however, FARS shows how activity in prefrontal cortex can modulate this pathway on the basis of object recognition and task constraints. For example, should one grasp a mug by the handle or the rim. The prefrontal cortex may tip the balance one way or the other, depending on whether the task at hand is to drink from the mug or move it to clear the table. This makes the important point that the selection of an appropriate action is based on multiple sources, an idea developed in the ACQ model (Section 2.3.3).

2.3.2 The mirror system

MNS, the Mirror Neuron System model of (Oztop and Arbib 2002), is based on the view that, when the monkey grasps an object, canonical neurons provide a premotor encoding of the type of grasp employed. The grasp will conform to one of the *affordances* of the object (e.g., the shape of one of the graspable parts of the object). MNS then provides a learning mechanism, which trains potential mirror neurons to associate visual input encoding the trajectory of a hand relative to an observed object with the canonical neuron encoding of that grasp. Since the visual input encodes hand movement relative to the object (or more specifically, to one of the affordances of the object), rather than retinotopically, the trained system is then able to recognize the actions of others because, even though the view of the self's hand is very different from one's view of the other's hand, the "object out" view of how the hand is positioned relative to the object's affordances remains the same in both cases. Thus, even in the absence of canonical neuron activity, there will be activation of mirror neurons associated with the observed object-centered trajectory of the other's behavior. MNS provides a bottom-up model of how the brain could learn to recognize movement as part of an action in an object context. In Section 4 we will present work toward recognizing a movement as meaningful in itself, even when there is no object with suggested affordances present in the current scene.

The MNS model utilized a feed-forward neural network with one hidden layer, which was trained using backpropagation. Such a network required an unnatural recoding of its input from the temporal to spatial domain. (Bonaiuto *et al.* 2007) developed a model, MNS2, which 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. This was a Jordan-type recurrent network trained using backpropagation through time (Werbos 1990).

As previously mentioned, the mirror neurons in the macaque can respond to observation of a grasp directed toward a recently observed, but currently occluded object (Umiltá *et al.* 2001). MNS2 incorporates working memory and dynamic remapping components, 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 (in simulation) 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.

Note that this ability to use the working memory of the location and affordances of an object is not adequate to support pantomime where the observer must, in general, infer the nature of the object from a movement similar to one which would be directed to such an object. For this reason, pantomime does not occur until a later stage than action recognition in the evolutionary process hypothesized in the Mirror System Hypothesis.

MNS2 further addresses data on “audiovisual” mirror neurons (Kohler *et al.* 2002). (Bonaiuto *et al.* 2007) associate each sound with a distinct pattern of activity 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.

2.3.3 Sequential action selection

While the FARS model focuses on reaching and grasping, a full model of imitation requires a more general account of the reproduction of sequences of actions. Before addressing the issue of sequence imitation, however, we describe recent modeling of flexible sequence production called augmented competitive queuing (ACQ). This model includes a mirror system for learning based on self-observation.

A classical model of sequence production is *competitive queuing* (CQ) (Bullock and Rhodes 2003; Houghton and Hartley 1995), which converts a spatial representation of a sequence into a temporal pattern of execution. The basic structure of the model is a three layer neural network. The first layer of the CQ network contains a single unit for each stored sequence. The next two layers each have units corresponding to all the basic actions from which the sequences are composed. Activation of a unit in the first, sequence storage layer, in turn activates a parallel representation in the parallel planning layer. Each unit in the parallel planning layer projects to a corresponding unit in the third layer – the competitive choice layer. This layer implements a winner-take-all process in which the most active element is selected for execution by temporarily inhibiting the other, less active elements. The winning unit thereafter inhibits its corresponding unit in the parallel planning layer (inhibition of return), removing it from the competition to determine subsequent actions. In this manner the spatial sequence representation in the parallel planning layer is converted into a temporal sequence of firing units in the competitive choice layer, in such a way that the higher the weight of the projection to its unit from the sequence storage layer unit, and thus the higher its activity in the planning layer, the earlier the corresponding action occurs in the sequence.

A surprising example – a cat reaching for food that is in a glass tube – shows the power of flexible scheduling of action. (Alstermark *et al.* 1981) lesioned the spinal cord of cats in order to determine the role of propriospinal neurons in forelimb movements. In particular, lesions in spinal segment C5 of the cortico- and rubrospinal tracts interfered with the cat's ability to grasp the food, but not to reach for it. The experimental setup consisted of a piece of food placed in a horizontal tube facing the cat. In order to eat the food, the cat was required to reach its forelimb into the tube, grasp the food with its paw, and bring the food to its mouth. Not reported in the paper is the account (B. Alstermark, personal communication to last author, 1990) that after the lesion, the cat would reach inside the tube, and repeatedly attempt to grasp the food and fail. However, these repeatedly unsuccessful grasp attempts would eventually succeed in displacing the food from the tube by a raking movement, and the cat would then bend its head down, grasp the food from the ground with its jaws and eat it. After only two or three trials, the cat began to rake the food out of the tube, a more efficient process than random displacement by failed grasps. The fact that after lesioning it took only a few trials for the cat to develop a successful motor program suggests that a form of learning was involved that takes place on a faster time scale than classical models of motor learning.

On the basis of these and other considerations, we have developed (Bonaiuto and Arbib, submitted for publication) a form of *augmented competitive queuing* (ACQ). A key difference from "classical" CQ is that the activation levels of motor program elements are dynamically computed in each "time step" rather than being completely specified before sequence execution and there is no inhibition of return. ACQ is based on three principles:

(1) Behavior emerges dynamically via the cooperation and competition of interacting perceptual and motor schemas.

(2) Motor schema activation is determined by a *priority signal*, computed in the parallel planning layer, that increases with both executability and desirability. *Executability* is determined by available affordances in the environment and the estimated probability of an action's success. *Desirability* represents the estimated value of an action in leading to reward, depends on current context and motivation, and is dynamically updated via reinforcement learning.

(3) An observation/execution matching (mirror) system may contribute to the rapid reorganization of motor programs in the face of disruption, when a known schema can be recognized as "filling the gap" for disrupted schemas.

This last point deserves special emphasis. It is common to think of the mirror system as encoding one's own intended actions and the observed actions of others. Here, we offer a radically new role for mirror neurons: the recognition of one's own unintended actions. This "mirror system for apparent actions", a new posited role for mirror neurons, comes into play because the actions to be reinforced within the current context are determined by *internal* and *external* recognition of self-generated actions. Internal action recognition is possible from an efference copy of the motor command just executed. External action recognition is determined by visual, auditory, tactile, and proprioceptive input. Typically, these signals coincide, but when they do not, multiple motor schemas can be reinforced. In the example of Alstermark's cat, we argue that the attempts to grasp the food that result in its displacement from the tube activate the mirror neurons for the action of raking the food from the tube, *even* *though the raking action was not intended*. The ACQ model demonstrates how this success can reinforce not only the action that was actually executed but also any action the mirror system recognizes during the course of that execution. The power of our model is that it provides a simple mechanism yielding a result that might otherwise seem to depend on high-level cognitive processes – supporting the flexible reorganization of coordinated control programs to achieve important goals despite changing circumstances.

We are further investigating the extension of ACQ to handle the learning of hierarchical motor programs. (Byrne 2003) describes the food processing techniques of gorillas (e.g., in gathering nettles and preparing them for "sting-free" eating) using flow diagrams to represent bimanually coordinated hierarchical motor programs. ACQ provides an alternative to flow diagram representations of actions by using competition between schemas based on dynamic ranking by priority. The flow diagrams describing gorilla-feeding behavior involve both competition between schemas and their cooperation in bimanual coordination. In determining how to do for Byrne's gorillas what we did for Alstermark's cat we are addressing the way in which complex behaviors introduce goals and subgoals. Desirability will then depend on the current subgoal rather than some overarching goal in the same way that secondary reinforcers may displace primary reinforcers in guiding animal behavior.

2.4 Complex imitation

The main mechanisms of complex imitation such as hierarchical action decomposition and reconstruction have been reported (Buccino *et al.* 2004), e.g., for imitation learning of guitar chords by musically naive participants. (Vogt *et al.* 2005) repeat the experiment and control for previous experience and learning during the experiment. They conclude that imitation learning of new motor patterns may work by decomposing the observed actions into elementary motor acts that activate and tune corresponding motor representations via mirror mechanisms. However, as we move from music to both praxic and communicative actions, we stress with (Wohlschläger *et al.* 2003) the notion of *goal-directed imitation* based on perceiving an action in terms of a (possibly incomplete, possibly erroneous) hierarchical structuring of goals and subgoals; see also (Buccino *et al.* 2004). Complex imitation thus involves "parsing" a novel action into a structured composite of pieces that achieve various subgoals, and finding out how these can be matched and melded together by variations on familiar acts. These representations are then recombined according to the observed model by the prefrontal cortex.

Importantly, novel actions can be acquired as skills through successive approximation over repeated trials, which may be needed to approach a finer-scaled decomposition of the action or to tune motor schemas to match less familiar pieces. However this last observation applies both to the statistical learning of hierarchical structure as in program level imitation by great apes (Byrne 2003) and the one-trial extraction of goal/subgoal structure in complex imitation, which is unique to the hominid line. It is likely that complex imitation requires some sort of symbolic representation overlaid on the lower-level representation. This is supported by studies of the effects of symbolic coding of observed actions on imitative accuracy (Bandura and Jeffery 1973; Carroll and Bandura 1987, 1990).

3. Gesture – embodied communication meets praxic action

At first it might be thought that the notion of goal-directed complex imitation is suited only for praxic, object-directed action where each movement has the goal of either positioning the end-effector or using it to change the state of some object. But what of intransitive movements like communicative gestures? Here the goals are more abstract. In communication, each word or gesture is part of achieving a communicative intention or (sub)goal; cf. (Arbib 2006). Even in the case of guitar playing, the movement has meaning only as part of a chord within a defined musical system, or serves to extend the system. However, humans are also able to imitate intransitive "meaningless" movements that are neither directed towards objects nor are part of a socially structured system of gestures. We thus first return to the study of primates as offered by the mirror system hypothesis to show why it is plausible to see the ability to reproduce "meaningless" movements as a concomitant of the ability for complex imitation.

3.1 The utility of "meaningless" movements

We have characterized complex imitation as involving "parsing" a novel action into a structured composite of pieces that achieve various subgoals, and finding out how these can be matched and melded together by variations on familiar acts. We further stressed that novel actions can be acquired as skills through successive approximation over repeated trials, which may be needed to approach a finer-scaled decomposition of the action or to tune motor schemas to match less familiar pieces.

The point here is that in mastering a novel action through complex imitation, we might first see a movement that achieves a particular subgoal, yet match it to one that does not achieve that subgoal. For example, we might at first think that piece A can be secured to piece B by inserting a rod on A into a hole on B, only to find that when we try this, the two pieces fall apart. After several trials, our failure draws our attention to what would hitherto have seemed to be a "meaningless" movement – a twisting of A that occurs after the insertion of the rod into the hole in B. We thus acquire a new subaction – insert plus twist – which then becomes part of our successful overall action. The suggestion is that the ability to observe and reproduce "meaningless" actions was necessary for successful imitation of complex actions that contained subactions, which could at best be approximated by an action already in the imitator's repertoire. Of course, once the novel action becomes incorporated as part of the means to achieve a subgoal, it is no longer meaningless, but instead becomes tuned as skill in achieving that subgoal.

We have already noted that apes could acquire a limited range of communicative gestures through ontogenetic ritualization (and, presumably, forms of social learning that build upon it). This, like simple imitation, requires long processes of learning, made even harder by the absence of physical goals. We suggested earlier that Stage S5 of the Mirror System Hypothesis – the emergence of protosign – involved *pantomime*, first of grasping and manual actions and then of non-manual actions, and *conventional gestures* that simplify, disambiguate or extend pantomime. In each case, we see the importance of recognizing a motion not as part of achieving a known *praxic goal* but rather as serving the *communicative goal* of stimulating mental retrieval of a physical goal, either through a more-or-less direct association of movements as in the case of pantomime, or more indirectly as in the case of other kinds of gestures (of which conventional ones are a special case).

In summary, while the analysis of an observed action can result in hierarchical structures of task-level subgoals that capture the movements' outcome, such as making a clapping noise irrespective of the body parts involved, or getting an object to a new location irrespective of which trajectory is followed, the analysis might also include "meaningless movements" defined by "motor level goals", which involve the movements' significant spatial or kinematic parameters, such as moving the hand to the nose or along a circular trajectory, but do not relate directly to external objects in the way that transitive actions do (cf. (Hermsdörfer *et al.* 2001)). We noted earlier that pantomime supported the emergence of protosign as a system of conventionalized signs, but is not itself part of protosign. Here we note that when protosign contains a relatively small number of signs, each may have been producible and recognizable as a separate entity. However, as the number of signs (whether vocal or manual or both) multiplied, keeping them distinct would have been increasingly onerous. The resulting adaptation seems to have been the development of what for speech is called “phonology” – the emergence of a small but stable set of meaningless units, from which meaningful units could be constructed (e.g., the duality of patterning between phonemes and words in spoken language – see (Studdert-Kennedy 2002) for how this may relate to mirror neurons). In all well-developed modern sign languages (which are fully expressive human languages, not protosign systems), each sign is composed from a restricted set of hand shapes, trajectories and start- and end-points. In pantomime, individual mimes may seize on quite different aspects of the event or object they are trying to convey, and may express those aspects idiosyncratically (just imagine the different ways you could try to convey the notion of “tree” in pantomime). This means that ambiguity can only be avoided by increasing the detail of the pantomime – or by only using a conventionalized form of the pantomime that is agreed on by the community and is carefully structured to minimize confusion with other pantomimes. This can mark the transition from pantomime to protosign, an important step toward the integration of hands, face and voice in protolanguage, prior to the (possibly cultural) evolution of an ever enlarging lexicon and ever increasing subtlety of syntax that marks the transition to language.

3.2 Imitating gestures

We have already noted that humans are able to imitate “meaningless”, intransitive movements, and that this ability may naturally accompany an ability for complex imitation. The question now is how this capacity is employed in social interactions to facilitate understanding of the various communicative body movements that other individuals perform, and how it can foster the establishment of empathic couplings that may enable inter-personal alignment and coordination. In fact, mimicry of expressive, referential motor acts is found in natural dialogue (Kimbara 2006). In the following, we will concentrate on a particular kind of hand movements in natural conversations, intuitively performed by speakers in order to materialize their communicative ideas and attended to by listeners in trying to pick up these ideas: iconic gestures.

Inspired by the categories of (Peirce 1960), who divided signs into Icons, Indexes, and Symbols[2], gestures are often considered as semiotic signs which may be distinguished according to the relation that holds between their overt form (the signifier) and the entity they refer to (the signified) (Kendon 2004; McNeill 1992, 2005). *Symbolic* gestures like those in sign languages or “emblems” like the victory sign hand shape have a clear-cut meaning which is fixed, determined by convention, arbitrarily associated with their form, and mostly independent of the accompanying speech. The majority of gestures in daily use, however, belong to the class of “gesticulations” (sometimes known as co-speech gestures), movements that do not have an unequivocal meaning and appear intimately related to the content of speech, with which they are temporally coordinated. Gesticulations can be subdivied into deictics and iconics, while keeping in mind that these should be considered dimensions rather than disjunctive categories (McNeill 2005). *Indexicality* or *deixis* of a gesture concerns the degree to which it directly refers to an entity in the extra-gestural context, e.g. by pointing to a present or imagined object in front of the speaker. *Iconicity* concerns the degree to which the gesture refers by virtue of its resemblance with the signified, i.e. by creating a gestural “picture” of its referent.

Here, we focus on the iconic aspect of gestures. Iconic gestures can comprise every possible posture or movement, as suited to create a gestural depiction of whatever the speaker

[2] See e.g. http://plato.stanford.edu/entries/peirce-semiotic

wants to refer to. They are often invented and hence novel to the observer. (McNeill 1992) deems their semiotic nature global and synthetic, that is, the meaning of the parts of the gesture are determined by the meaning of the whole, and several distinct meanings are merged into one gestural sign. Several researchers (Kendon 2004; Streeck to appear) have subdivided iconic gestures according to the depiction strategies or "practices" by which people describe the world with their hands. These categorizations are amenable to the embodied communication approach we are taking here, and we coarsely unify them into three classes of iconic gesture:

- *Pantomime* (or *enacting*): the hands represent themselves or other effectors in the course of action (e.g., the flapping wings of a bird).
- *Depicting*: the hands draw a 2D outline or sculpture (e.g. see Figure 3) a 3D shape in the air.
- *Modeling*: the hands or other body parts themselves act as tokens for the depicted entities.

The point here is that these different kinds of gestures, whose movements are more or less and in different ways abstracted away from transitive action, may require different kinds or degrees of involvement of the neural circuits that underlie simple and complex imitation as discussed above. A considerable body of research (examples given below) on patients who suffer from apraxia has already shown that certain other features of a gesture may influence how it is perceived, processed, and imitated. Apraxia is a disorder of voluntary, skilled movement due to a disorder of higher-level perceptual, cognitive, or motor systems. Here, "ideomotor apraxia" is particularly relevant, which classically refers to a derailment of the performance of generally well-conceived acts mainly due to disturbances at the stages of retrieval of motor representations. Apraxics are known to have specific deficits in the recognition, production, or imitation of three types of gestures: object-directed pantomimes of praxic action, symbolic gestures, and meaningless, non-referential postures or movements. Note, however, that these tasks (though mastered by healthy participants) are highly artificial situations where gestures are produced without a natural conversational context.

Several dissociations are found in the symptoms of ideomotor apraxics. (Morlaas 1928) first reported different abilities for transitive use of objects and intransitive gestures. This was corroborated in later experiments by (Cubelli *et al.* 2000) who conclude that transitive action and symbolic gestures rest upon distinct semantic and motor representations that can be selectively impaired. Other types of dissociations concern the nature and novelty of the stimulus that evoked the patient's response. Some patients can perform a symbolic gesture to verbal command (e.g. waving), but cannot recognize or imitate it upon immediate demonstration. Others can comprehend and name a demonstrated gesture, but cannot produce it themselves. Some are impaired in imitation of novel "meaningless" gestures, with normal performance on meaningful gestures (a disease called "visuo-imitative apraxia"). Others are not able to imitate a familiar gesture but can reproduce a novel movement (Bartolo *et al.* 2001; Goldenberg and Hagmann 1997). That is, the ability to copy a gestural movement pattern appears double-dissociated from the ability to get the meaning of a pantomimed action. We call the first ability *low-level* imitation*,* to distinguish it from imitation based on recognition and "replay" of a goal-directed action.

FIGURE 1 ABOUT HERE

A variety of cognitive models were proposed which seek to explain these dissociations by assuming a multi-stage translation process, with different stages specific to the imitation of novel or familiar gestures. The classical *dual route imitation learning* model (Rothi *et al.*

1991; see Figure 1, right part) posits a *direct route* for the imitation of meaningless gestures. It is assumed to convert *de novo* a visual representation of limb motion into intermediate postures or motions for subsequent execution. Damage to this route impairs novel gesture imitation, but does not affect the ability to imitate or name a seen meaningful gesture. By contrast, the model assumes that all meaningful gestures, whether or not they are object-directed, are recognized and then reconstructed for imitation via an *indirect route,* which involves two repositories of known actions, an *input* and an *output praxicon*. Damage to the input praxicon would specifically impair the ability to recognize/comprehend observed gestures, while damage to the output praxicon would exclusively impair the self-production of gestures. Both praxicons are connected to an *action semantic system* assumed to hold knowledge about the function of tools, the objects that participate in transitive actions, or the possible sequential organization of actions. The semantic system is connected to object recognition and verbal processes (Figure 1, left-hand side) and hence is involved in pantomime of actions to verbal command, the naming of displayed objects, and the imitation of meaningful transitive gestures.

The notion of a different mechanism underlying the low-level imitation of meaningless gestures from that of familiar actions has been supported by findings of segregated brain activation (Rothi *et al.* 1991). It concurs with neuropsychological reports to support the hypothesis that knowledge about the human body mediates between visual perception and motor execution. (Goldenberg and Hagmann 1997) suggested that the brain employs a common *body part coding* in representing novel gestures and body configurations in terms of spatial relationships between discrete body parts. This step permits recoding the detailed visual features of a gesture into a simpler form of a combination of familiar elements (Goldenberg and Karnath 2006), and it can provide a representational basis for formulating hierarchical motor level goals in imitation of intransitive gestures, as distinct from transitive gestures, which employ an object-centered representation. Errors in imitating a meaningless intransitive gesture could thus result from an inability to filter and convert the visual details of a demonstrated gesture to simpler body part codes, or to maintain this representation working memory until motor execution is completed.

FIGURE 2 ABOUT HERE

A body-centered movement representation may also be utilized in the indirect route. (Buxbaum *et al.* 2000) present evidence suggesting that both, imitation of familiar and novel movements require an intrinsic spatial coding of locations of body parts over time (a dynamic body model). In this model (see Figure 2, left), dynamic representations and procedures (“schemas”) are used to calculate movements and positions of body parts in all contexts. These modules are employed and supported by higher-level representations of learned gestures, as well as by more general knowledge of tool use which, in turn, connects to visual object input or speech input.

(Peigneux *et al.* 2004), who retain the distinction between a direct and indirect route, likewise assume that visual analysis provides a body part coding of gestures irrespective of the familiarity to the observer (see Figure 2, right). The direct route directly utilizes this body part coding and human body knowledge to implement imitation of novel gestures. The lexical route comprises one central praxicon with representations of the visual/shape and kinetic features of familiar gestures. These representations are supposed to be directly triggered by visual-gestural analysis, though no dedicated action recognition mechanism is featured, and to be transposable into innervatory patterns for the lower sensorimotor system. An action semantic system can trigger praxicon information for motor implementation of certain actions; conversely, information may pass from the praxicon to the semantic system for naming a familiar action.

We note, however, that apraxics who are unable to pantomime a particular action may nonetheless be able to perform it when they can recognize an object (and, in particular, the affordances) upon which the action is to be performed. This suggests complementary roles for body-based and object-based interactions in normal transitive actions. (Goldenberg *et al.* 2003) stress that, at the motor level, the action schema for object use is formulated based on adaptations to sensory data about the properties of the objects involved and the actual course of movement. Pantomimes differ crucially from the corresponding manipulative motor actions they depict as they, although they might rest on “motor memory” of execution of the pantomimed task, require the body-based emulation of the task on similar or different effectors, or rest on imagination of the affordances of an object to provide simulated inputs for use in the same motor schema as that which generates the pantomimed action.

As noted in our earlier discussion of Stage S5 of the mirror system hypothesis, from an evolutionary perspective it is important to distinguish pantomime as an ad hoc attempt to convey some action or object from conventionalized gestures, even if the latter may be related to pantomimes. Indeed, brain lesion studies (Corina *et al.* 1992; Marshall *et al.* 2004) of aphasics who were signers of American or British Sign Languages, respectively, show that there are lesions that preserve pantomiming while impairing the ability to use the meaningful signs of a sign language. By contrast, other researchers (Gallagher 2005; McNeill 2005) posit that even a pantomimed iconic gesture entirely serves a cognitive and communicative function. Consequently, such gestural movements are assumed to be controlled by a direct mapping of meaning onto space and motion through the linguistic/communicative system, without providing an exact account of this meaning and in how far it relates to action schemas and objects associated with a pantomime.

Another important aspect of iconic gestures is that they often come to be embedded in richly structured complexes, whose spatial and temporal arrangement can be novel (i.e., created by the speaker or seen by the listener for the first time) and employ different kinds of iconic gesture. For example, imagine someone performing a pantomime of drinking from a bottle, right after a depicting gesture (sculpting the bottle), while modeling the table the bottle is standing on with the other hand throughout. The overall intended meaning can only be extracted by analyzing the overall performance for its internal structure, interpreting each single gesture in this context, and combining these interpretations to form a cohesive representation of the entire action. That is, we see in gesture a form of structure and compositionality that does not follow standards of form, but is derived on the spot to convey the spatial or temporal features of the referent scene using gestures that appeal to different practices (sculpting, modeling, pantomime). Likewise, complex internal structures can be found even within a single iconic gesture, as demonstrated by the example in Figure 3. This gesture comprises three different, successively adopted postures with the palms facing each other along different axes. The internal structure, which does not impose an exact temporal order on the occurrence of the three expressive postures, derives from the need to cohesively depict the extents of the box in its three main spatial dimensions. This spatial arrangement needs to be taken into account when trying to understand or imitate the gesture. We have seen that complex imitation involves “parsing” a movement into a structure of (sub)goals, and we have noted that such (sub)goals can as well be communicative rather than praxic and object-related. Hence, we find in imitation of communicative iconic gestures the same basic mechanisms that we have assumed for imitation of grasping from an evolutionary perspective.

FIGURE 3 ABOUT HERE

Finally, *modeling* gestures may involve little if any influence of the observer's own motor system, beyond the recognition of the adoption of a salient posture. Such gestures can only be understood to the extent that the visual features of the employed body parts can be associated with spatial properties of the referents. For example, recognizing that the arm in its current posture is employed in a modeling gesture to represent a longish, straight object provides important clues to the interpretation of the gesture as representing a barrier.

4. Imitating gestures with virtual humans

When studying mechanisms of embodied communication in humans, one method of great heuristic value is the conception and evaluation of simulation models that need to figure in embodied interaction partners. One instance of that is the virtual human *Max*, developed at Bielefeld University's Artificial Intelligence Lab. Simulated with synthetic speech and a virtual body and face rendered in computer graphics, Max can engage in reciprocal interactions with real or other virtual humans. Here we focus on work that investigates how Max could, from observing and imitating others in such interactions, acquire and align with a human-like repertoire of expressive motor behaviors. A connected question is how Max's perceptual processes can internally be coupled with his active motor repertoire in order to facilitate empathic couplings and a faster, better understanding of the complex communicative goals behind a rather simple intransitive movement.

FIGURE 4 ABOUT HERE

4.1. Motor control and body-centered representation of movement

We start by describing how Max internally represents and processes the finely synchronized movements one can find in natural hand gesture. Max rests upon a "top-down" approach to motor control (shown in Figure 4), which starts from a compositional, body-centered specification of a gesture and breaks it down into a set of local controllers whose concurrent executions let emerge the desired movement (Kopp and Wachsmuth 2004). The initial specification, formulated in an XML language called MURML (for *Multimodal Utterance Representation Markup Language* (Kranstedt *et al.* 2002)), describes the significant morphological features of the gesture relative to a body-centered reference frame. A hand-arm configuration is defined in terms of four components: (1) the location of the wrist, specified in relation to the body by symbolic identifiers for the positions in the frontal, transversal, and saggital plane; (2) the configuration of the hand, compositionally described by the overall hand shape and modifiers for single finger flexions; (3) the direction at which the back of the hand is pointing; (4) the orientation of the palm (specified either absolutely as direction of the palm normal vector or relatively as rotation around the forearm). Such parameters are on a similar level of abstraction as those employed as input in modeling the macaque mirror system (Bonaiuto *et al.* 2007; Oztop and Arbib 2002) and the infant learning to grasp (Oztop *et al.* 2004), but with one critical exception: In those models, the emphasis is on transitive actions, and so parameters are specified relative to an object and its affordances, rather than the significant morphological features of the movement itself.

MURML builds on but extends a notation system for the German sign language of the deaf, *HamNoSys* (Prillwitz 1989). Commonly, sign language recognition is viewed as a pattern classification task. It is thus approached like speech recognition by employing techniques, which model the probabilities of the occurrence of fixed "phonemes" and their possible combinations for signs using techniques like Hidden Markov Models (Vogler and Metaxas 1998) and Artificial Neural Networks (as are used in the mirror system models referred to above). However, it is important to contrast the pre-defined vocabulary used in pattern classification approaches to sign language recognition, with the variety and idiosyncrasy of co-speech gestures that can be observed in everyday conversation. Treating a gesture as an emblem from a fixed repertoire would neglect the inner structures and possible commonalities of different gestures that could not be discerned in the all-or-none classification of the entire gesture. Therefore, gesture analysis and representation for Max is based on spatiotemporal features, such as hand configuration or relative motion, which MURML can represent.

Importantly, spontaneous iconic gestures tend to be highly variable and imprecise. MURML thus allows for a "least-commitment" representation that provides three ways to lay down only the required properties of a gesture. First, a gesture's spatiotemporal features are defined roughly, by using position and orientation symbols that correspond to a certain level of granularity. Secondly, only the features of the gesture's meaningful phase (often called "gesture stroke") are specified. And finally, the gesture can be underspecified by leaving features open that are not decisive for its expressive function. For example, the location of the hand in space may not be important for an isolated gesture, but may be crucial for signs in a sign language utterance.

As we have seen in the example above (Figure 3), a gesture is a complex combination of postures or sub-movements that make up its expressive phase, e.g., moving the hand upwards while keeping a fist. A sufficient representation for this must thus allow for decomposing the gesture into separate yet temporally coordinated features. We refer to these features as "movement constraints" and we take them to define the basic motor level goals that need to be fulfilled in order to constitute the motor act. MURML distinguishes between static (postural) constraints and dynamic (movement) constraints, which can be formulated for each of the aforementioned components. The internal structure of a gesture, and thus its composite of motor level subgoals, results from the relations that hold between these constraints. In our model, representations of the simultaneity, posteriority, repetition, and symmetry of movement constraints are indicated to compose a constraint tree that reflects the internal structure of the overall intransitive movement.

Max is based on an anthropometric kinematic skeleton (shown in part in Figure 4 at the right) that comprises 103 degrees of freedom in 57 joints, all subject to realistic motion limits. As mentioned above, Max's motor system takes a MURML definition as input and seeks to move his skeleton in such a way that the various constraints are reproduced reliably (e.g. fist hand shape while movement upwards). To this end, motor planning is decomposed as illustrated in Figure 4 into specialized motor subsystems (e.g. for the hands, wrists, and arms) that plan and instantiate *local motor programs* (LMPs), which then autonomously control motion within the corresponding limited set of degrees of freedom and over a designated period of time. For example, a single LMP can serve to move the hand (by accessing the shoulder and elbow joints), control a finger movement, orient the wrist, or lift the elbow. We distinguish different types of LMPs accordingly, as these scopes require different methods to control them, but stress that each LMP is universal in that it can be instantiated with a sequence of control parameters for the required target sub-movement. Such control parameters differ according to the kind of sub-movement described, e.g., arm movement is defined by a sequence of segments ("guiding strokes"), which define piece-wise the trajectory of the wrist through space (Kopp and Wachsmuth 2004). Together these parameter sequences form a motor plan, formulated in a body-centered representation that is also exploited for motor level imitation of gestures (as described below). Here, we stress that, although LMPs are closely related to the idea of distinct motor schemas, there are no fixed motion primitives like one controller for moving the hand to the right, or one for altering coupled joint angles in a rhythmic fashion – as is common in computational motor control as well (Schaal and Schweighofer 2005).

Multiple LMPs of various types can be queued and run in parallel within a more abstract *motor control program* (MCP). The exact combination and parametrization of LMPs is planned on the spot, depending on the structure and requirements of the single requested gesture. At execution time, LMPs activate and deactivate themselves as well as other LMPs, depending on feedback data about the current motion conditions. Note that we are dealing here with a simulated body that can provide immediate feedback, and so no forward models are required to derive predictions in order to compensate for feedback delay – although this could be easily imparted to the simulation. When an LMP becomes active it exerts immediate influence on some of the joints of Max's body. The overall gesture, and hence the overall solution to the control problem, then emerges from the concurrence of active LMPs.

4.2 Motor level imitation

In order to study the underlying mechanisms and functions of imitation with Max, we utilize a Virtual Reality setup in which a real human can engage in face-to-face, reciprocal interactions with a virtual human. In previous work (Kopp *et al.* 2004), the agent could mimic all the gestures demonstrated by the human, who was equipped with motion trackers and data gloves. Two separate systems were connected, one for gesture recognition and understanding (Sowa and Wachsmuth 2005) and one for gesture generation (Kopp and Wachsmuth 2004), linked via a gesture representation in MURML as described above. That is, the recognition module computed a MURML description after complete inspection of a demonstrated gesture, and the generation module processed it from scratch to reproduce the gesture. Together, these systems enabled a low-level, body-part-coding-based imitation of static, "meaningless" gestures, even when executed in a fast sequential manner. This procedure realized a form of direct route imitation. However, direct measurements of the human demonstrator were used and recognition and reproduction occurred in total separation. This was rather different from the suggested active role of the motor system in human perceptual processing (see Section 2), as well as from the approximate representation of a gesture that a human derives from visual observation of another human.

How can then the motor system be actively involved to support the perception of intransitive movements, e.g. to simulate automatic mimicry and immediate imitation of gesture on the basis of automatic response facilitation? In further work on motor level gesture imitation (Kopp and Graeser 2006), we follow cognitive models of apraxia in assuming different mechanisms for recognizing familiar and novel gestures (cf. Figures 1 and 2). Such models assume that familiar gestures are recognized via a lexical route that utilizes motor representations stored in a praxicon. That is, the motor system is more likely to play an integral role when a gesture is processed along the indirect route, whereas the direct route can run separate processes for visual input analysis and the mapping onto motor representations (but note the discussion below on route switching). In a related computational approach, (Demiris and Hayes 2002) distinguish between *passive* and *active imitation*. In passive imitation, the imitator runs a clear-cut "perceive-recognize-reproduce" cycle, with involvement of the imitator's own motor system only during the "reproduce" phase. In active imitation, the imitator's motor system is actively involved for recognition of familiar actions already during the perception process.

The classical approach by (Wolpert and Kawato 1998) to model active imitation, which we adopted here, is to form a forward model from each motor command under consideration, and use these models to derive predictions that can then be compared with the observed movement. Such an indirect route enables not only faster imitation, but also the automatic connection of an observed movement to the observer's own bodily and motor experiences as represented in the praxicon, and could thus support association of a gesture with occasions of one's own use of the gesture and hence its potential "meaning". If none of the forward models of the praxicon produces a good match, the direct route mechanisms may still determine motor level goals that enable an imitative response which may then, possibly over the course of multiple attempts in reciprocal interaction, lead to learning of the new behavior.

We have implemented the double route model of motor level gesture imitation (see Figure 5) and applied to a scenario where two virtual humanoid agents, *Max* and *Moritz*[3], meet each other. Moritz acts as the demonstrator and executes predefined or random movements, whereas Max takes on the role of the learner and imitator that initially has only little or even no motor knowledge stored in his praxicon. Max rests on the motor control model we have described in the previous section. Further, Max is equipped with a perception system comprising a view sensor for simulating his field of view, an ultra short-term sensory buffer for compensating sensory drop-outs, and a perceptual buffer that maintains a set of percepts as result of current sensory stimuli. When Max initially sees Moritz, he perceives and recognizes Moritz's hands, wrists, and elbows, such that percepts are created which henceforth describe the current positions of the corresponding body parts. Additionally, the hand percepts directly provide information about finger postures.

Above we noted that the brain seems to code observed biological movements in a body-centered representation. Likewise, the perceptuo-motor system of Max needs to transform the initially viewpoint-centered coordinates of Moritz's hands into a form that encodes their position in relation to first Moritz's and then his own body. This presupposes identifying correspondences between single parts of the demonstrator's and imitator's body – a hard problem that the primate nervous system also has to face, especially when coping with different body sizes and proportions. Although approaches to learning such body mappings exist in robotics, e.g. (Breazeal et al. 2005), we so far by-pass this problem with our virtual humans in that they have the same body proportions, and employ the same body-centered representation in both agents. All perceptual information (wrist position and orientation, finger posture) is then fed into a visual working memory that maintains a mental image of the currently observed movement.

FIGURE 5 ABOUT HERE

Figure 5 shows an outline of the double route motor level imitation; Figure 6 shows snapshots from an example interaction between Max and Moritz. Visual gesture input is acquired as described and stored as percepts in chronological order in visual working memory. The indirect route comprises an action recognition system that hinges on Max's praxicon; the direct route consists of inverse models that can break down the memory trace of a demonstrated movement into simple components. Both routes operate on the same body-centered motor representation.

The central component of the indirect route is the praxicon, which is organized as a graph as proposed, e.g., in (Buchsbaum and Blumberg 2005) or (Johnson and Demiris 2005). In the Buchsbaum and Blumberg model, however, nodes directly describe body postures and the edges interpolations between them. Max's praxicon contains as nodes representation of states of the own motor system and body, but the edges represent motor commands that cause the motor system to undergo a change from one configuration to another. Motor level goals can here be seen as the combination of an edge and its end node, representing the desired new state together with decisive features of the transition leading into it, and these goals can be queued along their organization in the graph. We thus model aspects of goal-based imitation

[3] Reminiscent of *Max and Moritz (A Story of Seven Boyish Pranks)*, a blackly humorous tale (in German), written and illustrated by Wilhelm Busch and published in 1865.

already at the motor level, but note that this is somewhat simpler than the hierarchy of subgoals and associated actions posited in our earlier discussion of complex imitation.

The motor command graph serves as a praxicon entry on which processes of recognizing, imitating, and predicting the action of other's operate. As illustrated in Figure 5, following classical approaches, forward models are employed during observation of a demonstrated movement to selectively activate motor codes that can successfully predict an observed behavior. The graph structure ensures that forward models are created always dependent on the motor context, instead of testing all motor acts in the repertoire every time. All motor commands that lead away from the currently active node (the state the agent is in), hence seem applicable, are turned into forward models and computed in parallel to derive predictions of the possible future courses of movement. These predictions are compared to the actual observations until only one motor command with a prediction error below a predefined threshold prevails. This winning edge can then be used to construct a motor program in order to initiate an immediate imitation (response facilitation), thus entering a reciprocal interaction well before the demonstration has ended as illustrated in Figure 6, left and middle.

Direct route imitation does not involve a praxicon, but is based on the idea that every movement needs to be composed out of simpler elements that can be parameterized to some extent. Consequently, two kinds of inverse models are employed along the direct route. The first inverse model accomplishes segmentation, i.e. it is an inverse model for the sequencing of motor commands, based on general kinematic properties of human hand-arm movement: the velocity profile is searched for local minima that represent significant drops of movement speed, and the directions of movement are grouped into clusters based on their similarity with an average movement direction and the overall curvature of the trajectory. The second inverse model is applied to each segment and determines the command parameters that are most likely to reproduce it.

The underlying assumption, here, is that we can even see the direct route as having a "vocabulary", e.g., of basic hand shapes and movements represented by a form of body part coding, so that in seeing a "meaningless" movement, we tend to decompose it in certain restricted ways (recall our discussion of gesture "phonology" above). However, this set of features is not so definite, and we do have the ability to extend it when our attention is drawn, e.g., to a mistake that occurs in imitation or interpretation of a gesture. We also note the distinction between execution of a fixed sequence in an intransitive gesture and the possibly repeated execution of some movements to achieve (sub)goals in a transitive action.

The finally derived motor command sequence is passed on as motor output for direct route imitation. Additionally, new edges are formed and added to the praxicon graph, leading from the currently active node via nodes for the end states of single segments (possibly newly inserted), to the final state of the demonstration. In that way, complex movements are translated into sequences of motor commands that are stored in the graph as paths spanning several edges and nodes. Parts of Moritz's movement that might be already known to Max are readily incorporated into the newly learned motor command sequence; see (Kopp and Graeser 2006).

FIGURE 6 ABOUT HERE

Generalizing from here we would assume that, when observing an action for imitation, be it transitive or intransitive, both routes are active at the same time and in a competitive manner. Since recognizing a familiar gesture should be faster than analyzing a novel gesture, the indirect route will win if the gesture is sufficiently familiar. In a recent study, (Tessari *et al.* 2007) found evidence that humans can strategically select which of the two routes to use, depending on the expected novelty or familiarity of the stimulus. In our double route model, the direct route cannot provide a full segmentation before the demonstration is nearly finished. While performing a lexical route imitation, on the other hand, Max continually monitors his movement and compares it with the observed one. In accord with our considerations above, we set deviation thresholds such that imitation is first and foremost goal-directed. That is, imitation is considered successful as long as a tolerance level is met with respect to the achievement of motor level goals, such as reaching a target position/configuration or reproducing the significant wrist trajectory features. Percepts of all movement segments successfully imitated in this sense are removed from working memory. If, however, the difference between the observed and the self-performed movement exceeds the threshold, the execution of the movement is interrupted (Figure 6, right) and Max returns to the last node at which concordance with the demonstrated was found. Since the chosen edge from this node on was not the right one, but rather the best fitting one, the now perceived movement must be new. In this case, imitation relies on the direct route. Thus our model leads us to obtain detailed insights into both, the different mechanisms involved in (here, motor level) imitation and how they may be computationally modeled in artificial humans.

5. Conclusions

The role of our body and the perceptuo-motor system in social interaction and communication are decisive aspects in Embodied Communication. One fundamental idea is that the observation of others' behavior may serve both to prime specific representational structures involved in the generation of our own communicative behavior, and to generate inter-personal "empathic couplings" that may coordinate and align interlocutors below the level of their intentional contributions. In this chapter we have elaborated on the role of imitation for embodied communication. While much existing work has addressed transitive, i.e., object-directed, actions, we took here a new perspective towards the perception, understanding, representation and reproduction of intransitive movements that predominate in communicative interactions.

We started by exploring mechanisms for the imitation of transitive actions and we traced an evolutionary path that leads from mirror neurons for praxic actions to forms of communication via increasingly complex forms of imitation. This underlined the crucial role of the capacity for complex imitation, providing the bridge from the limited gestural repertoires of apes to the open-ended protosign of early humans, as well as the notion of complex imitation as being goal-directed, i.e., based on "parsing" a novel action into a (possibly incomplete, possibly erroneous) hierarchical structuring of pieces that achieve various subgoals, and finding out how these can be approximated by variations on familiar acts. We also suggested why it is plausible to see the ability to reproduce "meaningless" movements as a concomitant of the ability for complex imitation.

The task at hand is to explore to what extent this account can be extended to intransitive bodily movements that are primarily communicative, as found with gestures – expressive movements of the hands – in natural dialogue. It has been hypothesized (Buccino *et al.* 2004; Vogt *et al.* 2005) that imitation and learning of (transitive) motor patterns may work by decomposing the observed actions into elementary motor acts that activate and tune corresponding motor representations via mirror mechanisms. Now, the human mirror system exceeds that of the macaque in that it also responds to observation of intransitive movement and even seems to code the form of the movement. It thus seems natural to assume that the way humans try to interpret and understand an intransitive, probably expressive movement parallels the ways proposed for their understanding of object-directed action. Furthermore, research on the specific impairments of gesture imitation in apraxia hints at possible architectural demands for this, namely, an indirect, lexical route vs. a direct route and a crucial role for a body-centered movement representation in both processes.

Carrying this view into communicative behavior, we note that each gesture is part of achieving a communicative goal. At first, the listener will observe "meaningless" movements, which cannot be (at least fully) interpreted. However, we suggest that the (possibly richly structured) movements are already conceivable in terms of structured motor level goals that involve a body coding of a movement's significant spatial or kinematic features. An extended human mirror system could help in recognizing the own primitive motor acts needed to realize them, and we allude here to the possible emergence of a small but stable set of meaningless movements units, from which meaningful units could be constructed (a "gesture phonology"). The fact that the capacity for recognizing intransitive movement appeared later in evolution than that for transitive actions might also suggest that transitive movements are simpler to encode, possibly because the coordinate frame may be fixed on the object, whereas intransitive movement involves body part coding of spatial configurations of moving limbs (which involves multiple moving coordinate frames). In any case, here, we see an instance of complex imitation already required for low-level imitation, but keep in mind that we generally have a simpler sequential goal structure than with complex action composites.

For communication to be successful, however, it is important to recognize motion as serving a communicative goal. (McNeill 2005) posited that gestures are entirely different from goal-directed actions such that, in speaking, vocal and gestural motor acts get directly activated in Broca's area and adjacent premotor areas by communicative goals and constrained by the meaning to be expressed. We note, however, the difference between praxic goals for transitive movements and communicative goals for gestures (Arbib 2006). One possible solution to bridge the gap between obviously goal-directed transitive actions and expressive intransitive movements (gestures) may be the different depiction strategies one can find in iconic gesture and which resemble praxic actions geared towards the creation of gestural images (driven by our communicative intentions). If we pantomime, we (more or less intentionally) carry out patterns of an action. If we depict, we carry out sculpting or drawing actions. If we model, we choose to employ a body part in a fundamentally different way, without performing any action at all. That is, a gesture could follow a communicative goal of stimulating retrieval of a physical goal – either through a direct association of movements (as in pantomime) or indirectly through a more or less metaphorical and impoverished use of practices (as in depicting or modeling).

The second line of research that we have described here is work on artificial humans, which aims for exploring models of communicative behavior and implementing them to advance human-machine interaction. We suggest that equipping computer systems with a properly modeled humanoid body – real or simulated – and endowing them with increasing abilities for imitation can be key to improve their abilities of engaging in natural communications with humans. From a practical point of view, learning a human-like repertoire of expressive motor behaviors from observing and imitating others promises to help overcome the behavior acquisition problems we face in present systems, that mostly use fixed repertoires of often tediously modeled actions or that even need to be tele-operated (Schaal 1999). Another important aspect is the lack of current systems to engage in what we have called here "empathic couplings". Mainly, this is due to the fact that input recognition and understanding commonly relies on models of "user action" that are (biologically implausibly) different and separated from the system's models of its own actions.

With the virtual human Max we started to explore these issues for expressive, intransitive movements, based on considerations of the more general mechanisms and functions of imitation laid out here. The double route imitation model we adopted provides a framework to account for both, imitation learning of novel motor acts (via the direct route) and the coupling of perceptual processes with the agent's own motor repertoire in a way that facilitates the recognition of familiar acts and fast imitative responses as in non-conscious mimicry. Yet, while appealing to the general notion of goal-based complex imitation, our model is basic in that it focuses on intentions in the sense of "motor level goals", not the communicative intention connected to the intransitive movement. Further work is to address how a generalization can take place to form more general gesture schemas, akin to our theoretically explored gesture practices, from motor level goals that are often roughly identical in some body-coded features and differ in a number of others.

Another issue for further research is how the indirect route can be extended to the meaning-based understanding and imitation of gestures, and this can be integrated into online processing of verbal input. The natural approach taken in models of world action understanding is to link primitive motor acts with representations of skilled, goal-directed action, connected with preconditions and achieved final goals, e.g. (Buchsbaum & Blumberg, 2005). This amounts generally to the capacity for complex goal-directed imitation, but, as noted earlier (and as is essential for pantomime), one must clarify here the relation between the object-centered representation of transitive actions and the more abstract body-centered representations of those intransitive gestures, which may in some sense be related to them. In former work, we have started to model a meaning-level imitation (Kopp *et al.* 2004) and understanding of shape-related gestural and verbal expressions, along a direct route approach that utilizes a dedicated formalism for representing visuo-spatial information (Sowa and Wachsmuth 2005). In such an imitation game, an observer would perceive a complete multimodal description, construct a representation of the conveyed visuo-spatial content, and then generate an imitation that recodes the same content, but may possibly employ alternative gestural and verbal forms.

Finally, we want to stress that Max is a humanoid emulated in computer graphics, rather than being physically embodied. Thus, we must face up to the paradox that what can be modeled quite straightforward for the virtual human – the reproduction of meaningless movements – is a latecomer in primate evolution, indeed being unique to real humans among extant primates. Max demonstrates how gestures may be implemented for a virtual human in a manner which facilitates human-computer interaction, although we have touched only the motor realization part of it here, and it provides a valuable test bed for theories on how gestures may both augment and supplement spoken language. However, where primate evolution has moved from embodiment to symbol as we have laid out, the inherent design of computers to process symbols has led us in the opposite direction in the development of virtual humans – from symbols towards effective simulation of the perception of human actions and their emulation in a virtual human.

The circle will be closed as these ideas come to be applied to artificial agents carrying out praxic actions in a realistically simulated or, even better, the physical world. Here, we may expect novel tasks conducted with both humanoid and non-humanoid robots to offer us fresh insights into embodiment which will help us understand what aspects of our human embodiment are indeed the traces of our primate history, and to what extent they point to a truly general theory of embodiment. Such a general theory will both aid our understanding of human communication and point to embodied patterns of communication between humans and future robots, as well as communication between robots whose effectors, actions and interactions differ dramatically from our own.

References

Alstermark B, Lundberg A, Norrsell U and Sybirska E (1981). Integration in descending motor pathways controlling the forelimb in the cat: 9. Differential behavioural defects after spinal cord lesions interrupting defined pathways from higher centres to motoneurones. *Experimental Brain Research,* **42**, 299-318.

Arbib MA (2005a). From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics (with commentaries and author's response). *Behavioral and Brain Sciences,* **28**, 105-67.

Arbib MA (2005b). Interweaving protosign and protospeech: Further developments beyond the mirror. *Interaction Studies: Social Behavior and Communication in Biological and Artificial Systems,* **6**, 145-71.

Arbib MA (2006). A sentence is to speech as what is to action? *Cortex*, **42**(4), 507-14.

Arbib MA and Rizzolatti G (1997). Neural expectations: a possible evolutionary path from manual skills to language. *Communication and Cognition***,** **29**, 393-424.

Bandura A and Jeffery RW (1973). Role of symbolic coding and rehearsal processes in observational learning. *Journal of Personality and Social Psychology,* **26**(1), 122-30.

Bartolo A, Cubelli R, Della Sala S, Drei C and Marchetti C (2001). Double dissociation between meaningful and meaningless gesture production in apraxia. *Cortex,* **37**, 696-99.

Bonaiuto J, Rosta E and Arbib MA (2007). Extending the mirror neuron system model, I: Audible actions and invisible grasps. *Biol Cybern,* **96,** 9-38.

Bonaiuto J and Arbib MA (Submitted for publication). What did I just do? A new role for mirror neurons.

Brass M, Bekkering H, Wohlschläger A and Prinz W (2000). Compatibility between observed and executed finger movements: comparing symbolic, spatial and imitative cues. *Brain and Cognition,* **44**, 124-43.

Brass M and Heyes C (2006). Grasping the difference: what apraxia can tell us about theories of imitation (letters response). *Trends In Cognitive Sciences,* **10**(3), 95-96.

Breazeal C, Buchsbaum D, Gray J, Gatenby D and Blumberg B (2005). Learning from and about others: towards using imitation to bootstrap the social understanding of others by robots. *Artificial Life*, **11**, 1-2.

Buccino G, Vogt S, Ritzl A *et al.* (2004). Neural circuits underlying imitation learning of hand actions: An event-related fMRI study. *Neuron,* **42**, 323-33.

Buchsbaum D and Blumberg B (2005). Imitation as a first step to social learning in synthetic characters: A graph-based approach. In *Symposium on Computer Animation*, pp. 9-18. ACM Press.

Bullock D and Rhodes BJ (2003). Competitive queuing for planning and serial performance. In MA Arbib, ed. *The Handbook of Brain Theory and Neural Networks, Second Edition*, pp. 241-44. Cambridge, MA, A Bradford Book/The MIT Press.

Buxbaum LJ, Giovannetti T and Libon D (2000). The role of the dynamic body schema in praxis: evidence from primary progressive apraxia. *Brain Cogn,* **44**(2), 166-91.

Byrne RW (2003). Imitation as behavior parsing. *Philosophical Transactions of the Royal Society of London (B),* **358**, 529-36.

Byrne RW and Byrne JME (1993). Complex leaf-gathering skills of mountain gorillas: Variability and standardization. *American Journal of Primatology,* **31**, 241-61.

Caro TM and Hauser MD (1992). Is there teaching in nonhuman animals? *The Quarterly Review of Biology,* **67**, 151-74.

Carroll WR and Bandura A (1987). Translating cognition into action: The role of visual guidance in observational learning. *Journal of Motor Behavior,* **19**(3), 385-98.

Carroll WR and Bandura A (1990). Representational guidance of action production in observational learning: A causal analysis. *Journal of Motor Behavior,* **22**(1), 85-97.

Chartrand TL and Bargh JA (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality & Social Psychology,* **76**(6), 893-910.

Corina DP, Poizner H, Bellugi U, Feinberg T, Dowd D and O'Grady-Batch L (1992). Dissociation between linguistic and nonlinguistic gestural systems: a case for compositionality. *Brain and Language,* **43**(3), 414-47.

Corp N and Byrne RW (2002). Ontogeny of manual skill in wild chimpanzees: evidence from feeding on the fruit of saba florida. *Behavior,* **139**, 137-68.

Craighero L, Bello A, Fadiga L and Rizzolatti G (2002). Hand action preparation influences the responses to hand pictures. *Neuropsychologia,* **40**, 492-502.

Cubelli R, Marchetti C, Boscolo G and Della Sala S (2000). Cognition in action: testing a model of limb apraxia. *Brain Cogn,* **44**(2), 144-65.

Demiris Y and Hayes G (2002). Imitation as a dual-route process featuring predictive and learning components: a biologically-plausible computational model. In K Dautenhahn and C Nehaniv, eds. *Imitation in Animals and Artifacts*. Cambridge, MA, The MIT Press.

Fagg AH and Arbib MA (1998). Modeling parietal-premotor interactions in primate control of grasping. *Neural Netw,* **11**(7-8), 1277-303.

Ferrari PF, Gallese V, Rizzolatti G and Fogassi L (2003). Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. *Eur J Neurosci*, ***17***(8), 1703-14.

Ferrari PF, Rozzi S and Fogassi L (2005). Mirror neurons responding to observation of actions made with tools in monkey ventral premotor cortex. *J Cogn Neurosci*, **17**(2), 212-26.

Gallagher S (2005). *How the body shapes the mind.* Oxford, Oxford University Press.

Gallese V and Goldman A (1998). Mirror neurons and the simulation theory of mind-reading. *Trends Cognit. Sci.,* **2**, 493-501.

Gibson JJ (1979). *The ecological approach to visual perception*. Boston, Houghton Mifflin.

Goldenberg G and Hagmann S (1997). The meaning of meaningless gestures: A study of visuo-imitative apraxia. *Neuropsychologia* **35**(3), 333-41.

Goldenberg G, Hartmann K and Schlott I (2003). Defective pantomime of object use in left brain damage: apraxia or asymbolia? *Neuropsychologia,* **41**, 1565-73.

Goldenberg G and Karnath HO (2006). The neural basis of imitation is body part specific. *J Neurosci,* **26**, 6282-87.

Hermsdörfer J, Goldenberg G, Wachsmuth C *et al.* (2001). Cortical correlates of gesture processing: Clues to the cerebral mechanisms underlying apraxia during the imitation of meaningless gestures. *NeuroImage,* **14**, 149-61.

Houghton G and Hartley T (1995). Parallel models of serial behavior: Lashley revisited. *Psyche,* **2**(25).

Jacob P and Jeannerod M (2005). The motor theory of social cognition: a critique. *Trends in Cognitive Sciences,* **9**, 21-25.

Johnson M and Demiris Y (2005). Hierarchies of coupled inverse and forward models for abstraction in robot planning, recognition and imitation. *Proc AISB symposium on imitation in animals and artifacts*, pp. 69-76.

Kendon A (2004). *Gesture: Visible action as utterance.* Cambridge, Cambridge Univ. Press.

Kimbara I (2006). On gestural mimicry. *Gesture,* **6**, 39-61.

Knoblich G and Prinz W (2005). Linking perception and action: An ideomotor approach. In HJ Freund, M Jeannerod, M Hallett and RC Leiguarda, eds. *Higher-order motor disorders*, pp. 79-104. Oxford, Oxford University Press.

Kohler E, Keysers C, Umiltà MA, Fogassi L, Gallese V and Rizzolatti G (2002). Hearing sounds, understanding actions: Action representation in mirror neurons. *Science,* **297**, 846-48.

Kopp S and Graeser O (2006). Imitation learning and response facilitation in embodied agents. In J Gratch, ed. *Intelligent Virtual Agents, LNAI 4133,* pp. 28-41. Berlin, Springer-Verlag.

Kopp S, Sowa T and Wachsmuth I (2004). Imitation games with an artificial agent: From mimicking to understanding shape-related iconic gestures. In A Camurri and G Volpe, eds.

Gesture-Based Communication in Human-Computer Interaction, LNAI 2915, pp. 436-47. Berlin, Springer-Verlag.

Kopp S and Wachsmuth I (2004). Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds,* **15**(1), 39-52.

Kranstedt A, Kopp S and Wachsmuth I (2002). MURML: A multimodal utterance representation markup language for conversational agents. Working notes AAMAS-02 Workshop *Embodied Conversational Agents-Let's Evaluate and Specify Them*.

Marshall J, Atkinson J, Smulovitch E, Thacker A and Woll B (2004). Aphasia in a user of British Sign Language: Dissociation between sign and gesture. *Cognitive Neuropsychology,* **21**, 537-54.

McNeill D (1992). *Hand and mind*. Chicago, The University of Chicago Press.

McNeill D (2005). *Gesture and thought.* Chicago, University of Chicago Press.

Meltzoff AN and Moore MK (1977). Imitation of facial and manual gestures by human neonates. *Science* **198**, 75-78.

Morlaas J (1928). *Contribution a l'etude de l'apraxie*. Paris, Legrand.

Myowa-Yamakoshi M and Matsuzawa T (1999). Factors influencing imitation of manipulatory actions in chimpanzees (*Pan troglodytes*). *J. Comp. Psychol,* **113**, 128-36.

Oztop E and Arbib MA (2002). Schema design and implementation of the grasp-related mirror neuron system. *Biol Cybern,* **87**(2), 116-40.

Oztop E, Bradley NS and Arbib MA (2004). Infant grasp learning: a computational model. *Exp Brain Res,* **158**(4), 480-503.

Peigneux P, Van der Linden M, Garraux G *et al.* (2004). Imaging a cognitive model of apraxia: The neural substrate of gesture-specific cognitive processes. *Human Brain Mapping,* **21**, 119-42.

Peirce CS (1960). Division of aigns. In C Hartshorne and P Weiss, eds. *Collected Papers of C.S. Peirce*. Cambridge, MA, Harvard University Press.

Pinker S and Bloom P (1990). Natural language and natural selection. *Behavioral and Brain Sciences,* **13**, 707-84.

Pollick AS and de Waal FBM (2007). Ape gestures and language evolution, *PNAS,* **104**(19), 8184-89.

Prillwitz S (1989). *HamNoSys. Version 2. Hamburger Notationssystem fűr Gebärdensprachen. Eine Einfűhrung*. SIGNUM-Verlag.

Rizzolatti G and Arbib MA (1998). Language within our grasp. *Trends in Neuroscience,* **21**(5), 188-94.

Rizzolatti G and Craighero L (2004). The mirror-neuron system. *Annu Rev Neurosci,* **27**, 169-92.

Rizzolatti G, Fadiga L, Gallese V Fogassi L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive brain research,* **3**, 131-41.

Rothi, LJ, Ochipa C and Heilman KM (1991). A cognitive neuropsychological model of limb praxis. *Cognitive Neuropsychology,* **8**, 443-58.

Schaal S (1999). Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences,* **3**, 233-42.

Schaal S and Schweighofer N (2005). Computational motor control in humans and robots, *Curr Opin Neurobiol*, **15**(6), 675-82.

Seyfarth RM, Cheney DL and Bergman TJ (2005). Primate social cognition and the origins of language. *Trends in Cognitive Sciences,* **9**(6), 264-66.

Sowa T and Wachsmuth I (2005). A model for the representation and processing of shape in coverbal iconic gestures. In K Opwis and IK Penner, eds. *Proc. of KogWis05*, pp. 183-88. Basel, Schwabe Verlag.

Sowa T and Wachsmuth I (2002). Interpretation of Shape-Related Iconic Gestures in Virtual Environments. In I Wachsmuth and T Sowa, eds. *Gesture and sign language in human-computer interaction, LNAI 2298*, pp. 21–33. Berlin, Springer-Verlag.

Stokoe WC (2001). *Language in hand: Why sign came before speech*. Washington, DC, Gallaudet University Press.

Streeck J (to appear). *Gesture: The manufacture of understanding*.

Studdert-Kennedy M (2002). Mirror neurons, vocal imitation and the evolution of particulate speech. In M Stamenov and V Gallese, eds. *Mirror neurons and the evolution of brain and language*, pp. 207-27. Amsterdam, John Benjamins.

Tessari A, Canessa N, Ukmar M and Rumiati RI (2007). Neurophysiological evidence for a strategic control of multiple routes in imitation. *Brain*, **130**, 1111-26.

Thorpe W (1956). *Learning and instinct in animals*. London, Methuen.

Tomasello M (1999). The Human adaptation for culture. *Annu. Rev. Anthropol.,* **28**, 509-29.

Tomasello M and Call J (1997). *Primate cognition*. New York, Oxford University Press.

Tomasello M, Kruger AC and Ratner HH (1993). Cultural learning. *Behavioral and Brain Sciences,* **16**, 495-552.

Umiltá MA, Kohler E, Gallese V *et al.* (2001). I know what you are doing: a neurophysiological study. *Neuron,* **31**, 155-65.

Vogler C and Metaxas D (1998). ASL recognition based on a coupling between HMMs and 3D motion analysis. In *Proc 6th IEEE Int'l Conf. Computer Vision*. IEEE Press.

Vogt S, Buccino G, Wohlschläger AM *et al.* (2005). The mirror neuron system and area 46 in the imitation of novel and practiced hand actions: an event-related fMRI study. In *Proc 23rd European Workshop on Cognitive Neuropsychology.* Bressanone.

Werbos PJ (1990). Backpropagation through time: what it does and how to do it. *Proc IEEE,* **78**(10), 1550-60.

Williams JH, Whiten A, Suddendorf T and Perrett, DI (2001). Imitation, mirror neurons and autism. *Neurosci Biobehav Rev,* **25**(4), 287-95.

Wilson M and Knoblich G (2005). The case for motor involvement in perceiving conspecifics. *Psychological Bulletin,* **131**(3), 460-73.

Wohlschläger A, Gattis M and Bekkering H (2003). Action generation and action perception in imitation: an instance of the ideomotor principle. *Phil Trans R Soc Lond,* **358**, 501-15.

Wolpert DM and Kawato M (1998). Multiple paired forward and inverse models for motor control. *Neural Networks,* **11**(7-8), 1317-29.

Captions

Figure 1. A dual route imitation learning model balancing language and praxis. We stress that the right-hand side should be augmented by an "action buffer", and emphasize the bidirectional link between lexicon and semantics. (Adapted from Rothi *et al.* 1991).

Figure 2. Cognitive models of upper limb apraxia, redrawn from Buxbaum et al. (2000; left) and Peigneux et al. (2004; right). Solid arrows represent the flow of information; dashed arrows indicate assumed connections, either not normally used or not yet experimentally confirmed.

Figure 3. A three-phase iconic gesture used to depict a three-dimensional box-like object (adapted from Sowa and Wachsmuth 2002).

Figure 4. *Max* is shown at right, showing his "skeleton" without the rendered body superimposed upon it. The block diagram at left provides an outline of Max's motor control system. Hand-arm movements are produced from body-centered feature representations by creating and tuning local motor programs that concurrently control necessary sub-movements.

Figure 5. Outline of the double route model of motor level gesture imitation.

Figure 6. Motor level imitation of intransitive gestures of Moritz (demonstrator, left) by Max (imitator, right). *Left*: Max imitates immediately upon recognizing the demonstrated gesture; *middle*: Max has successfully reproduced the gesture; *right*: Max interrupts and returns to the last correctly recognized state, if an identified motor act turns out to be incorrect.

Figures

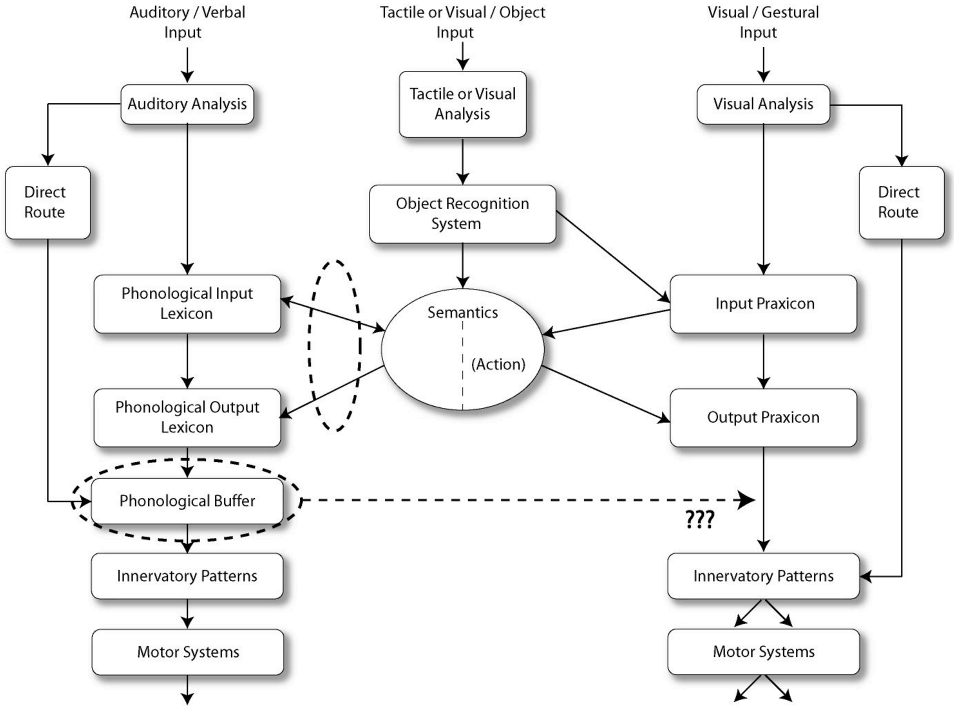

Fig. 1

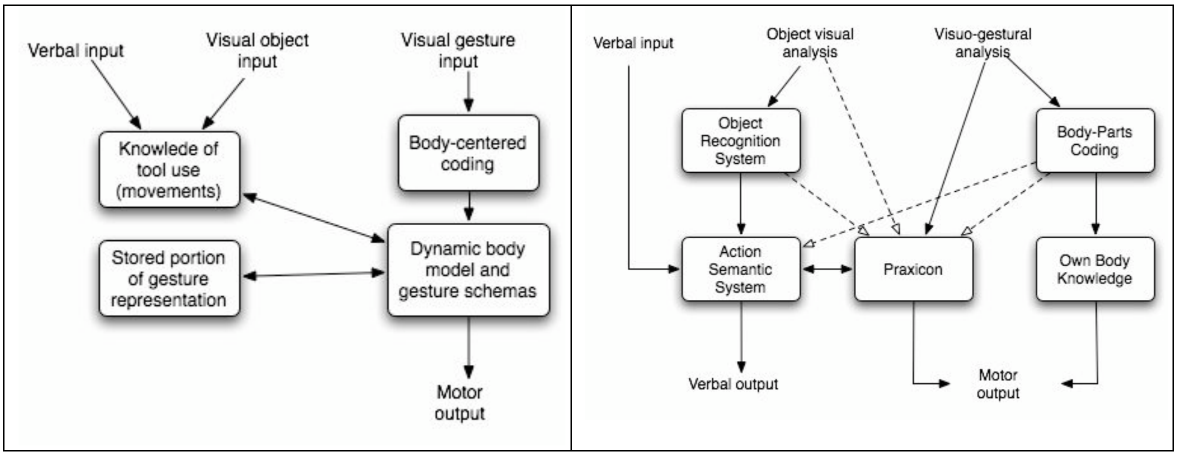

Fig. 2

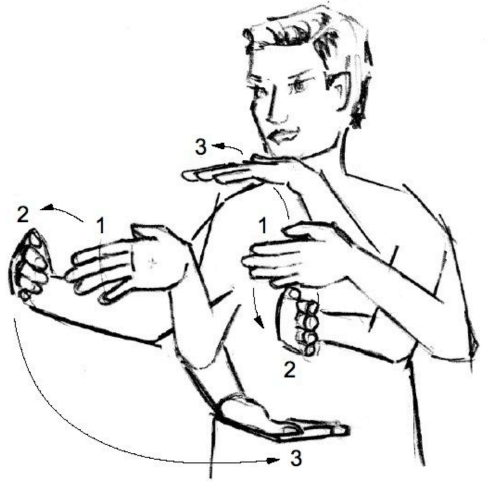
3
2
1
1
2
3

Fig. 3

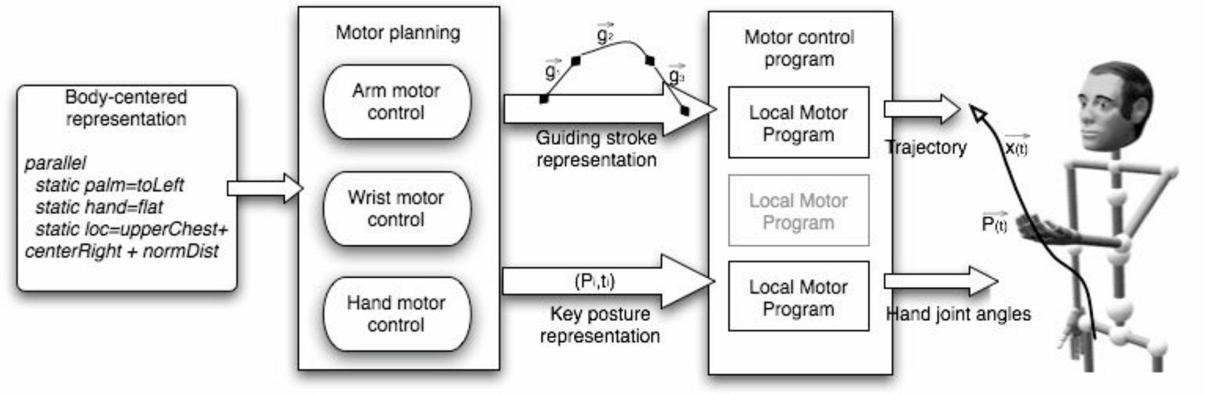
Body-centered representation
parallel
static palm=toLeft
static hand=flat
static loc=upperChest+
centerRight + normDist
Motor planning
Arm motor control
Wrist motor control
Hand motor control
$\vec{g}_1$
$\vec{g}_2$
$\vec{g}_3$
Guiding stroke representation
(P_i,t_i)
Key posture representation
Motor control program
Local Motor Program
Local Motor Program
Local Motor Program
Trajectory
$\vec{x}(t)$
$\vec{P}(t)$
Hand joint angles

Fig. 4

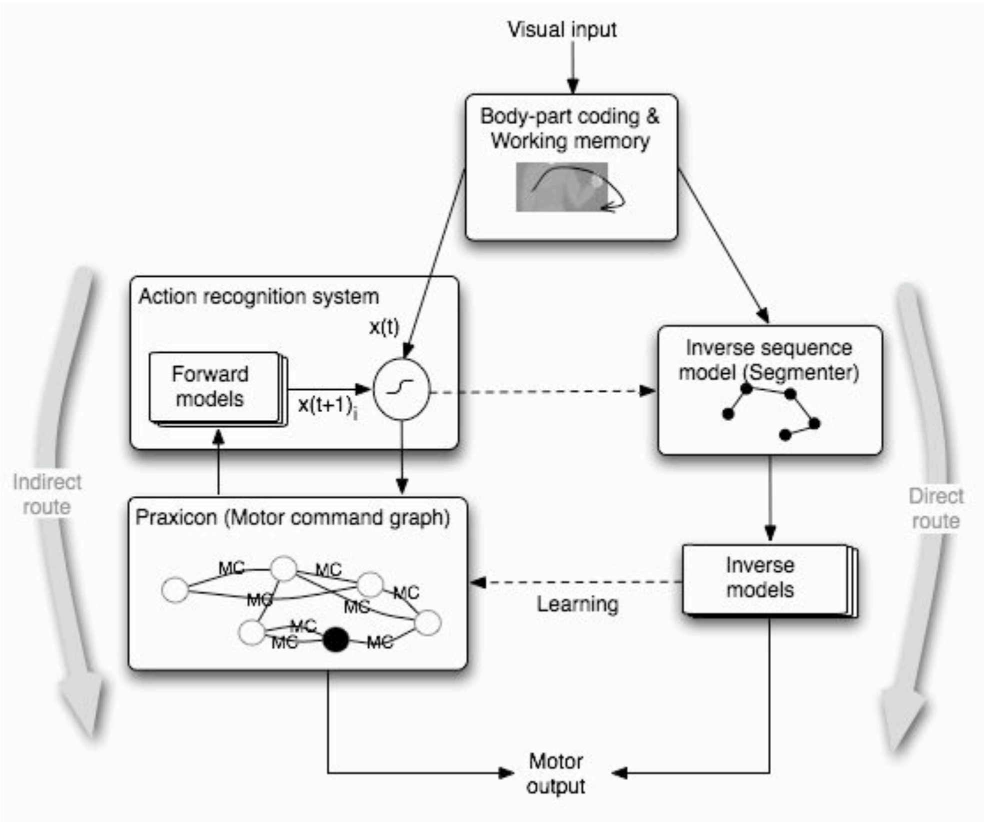
Visual input
Body-part coding & Working memory
Action recognition system
x(t)
Forward models
x(t+1)i
Inverse sequence model (Segmenter)
Indirect route
Direct route
Praxicon (Motor command graph)
MC
Inverse models
Learning
Motor output

Fig. 5

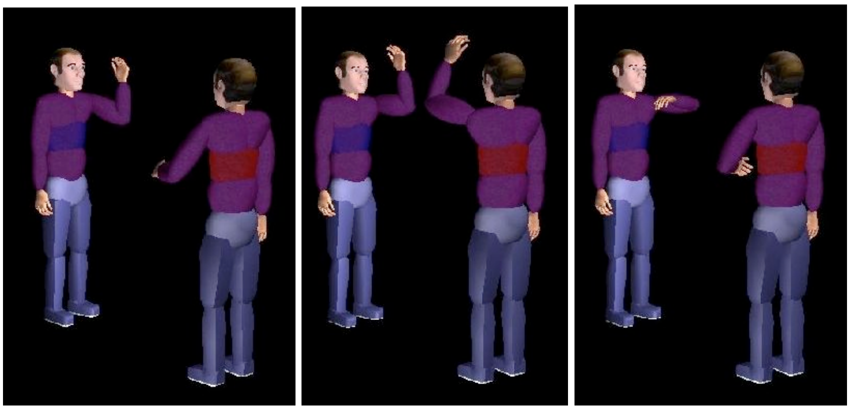

Fig. 6