Cogn Process
DOI 10.1007/s10339-009-0351-2

REVIEW

A model for production, perception, and acquisition of actions in face-to-face communication

Bernd J. Kröger · Stefan Kopp · Anja Lowit

Received: 31 July 2009 / Accepted: 13 November 2009

Abstract The concept of action as basic motor control unit for goal-directed movement behavior has been used primarily for private or non-communicative actions like walking, reaching, or grasping. In this paper, literature is reviewed indicating that this concept can also be used in all domains of face-to-face communication like speech, co-verbal facial expression, and co-verbal gesturing. Three domain-specific types of actions, i.e. speech actions, facial actions, and hand-arm actions, are defined in this paper and a model is proposed that elucidates the underlying biological mechanisms of action production, action perception, and action acquisition in all domains of face-to-face communication. This model can be used as theoretical framework for empirical analysis or simulation with embodied conversational agents, and thus for advanced human–computer interaction technologies.

Keywords Face-to-face communication · Speech · Co-verbal behavior · Action · Facial expression · Hand-arm gesture · Production · Motor behavior · Multimodal perception · Acquisition of action · Embodied conversational agents · Human–computer interaction

B. J. Kröger (✉)
Department of Phoniatrics, Pedaudiology,
and Communication Disorders, University Hospital Aachen
and RWTH Aachen University, Aachen, Germany
e-mail: bkroeger@ukaachen.de

S. Kopp
Sociable Agents Group, Center of Excellence
“Cognitive Interaction Technology”,
Bielefeld University, Bielefeld, Germany
e-mail: skopp@techfak.uni-bielefeld.de

A. Lowit
Speech and Language Therapy Division,
Department of Educational and Professional Studies,
University of Strathclyde, Glasgow, UK
e-mail: a.lowit@strath.ac.uk

Introduction

Actions are “goal-directed behaviors that usually involve movement” (Jahanshahi and Frith 1998, p. 483). Although one can also define intangible complex non-motor actions (e.g. long-term strategic goals such as conflict solving among different social groups), this paper will focus on *motor actions* involving movements of the body or of parts of the body (i.e. articulators), performed by a person (actor or task performer) in order to accomplish a specific *goal* or *task*. That is, actions are usually *motor specific* and they are always *goal-directed* (Shadmehr and Mussa-Ivaldi 1994; Sabes and Jordan 1997; Todorov 2004). It is possible to separate actions either as intentional and self-generated, i.e. arising from internal cognitive processes (*willed actions* Jahanshahi and Frith 1998, p. 483), or as arising from environmental stimuli (externally triggered voluntary or reflexive routine actions, Jahanshahi and Frith 1998, p. 483f; Latash 2008). Thus, willed actions form a subgroup of *voluntary actions*. Moreover, willed actions can be *private* or *communicative*. A private action can occur “in a private context, in mere fulfillment of the agent’s needs or goals or be part of a process involving communication between (…) individuals” (Jeannerod 1999, p. 1). An example for a private action is reaching for or grasping an object such as a cup on a table. An example for a communicative action is pointing with the goal of conveying a specific object location or direction to the interlocutor. The latter is a willed action as it is meant either to “display”,

Published online: 10 December 2009

i.e. intended to show, the indexical information or to even "signal", i.e. intended to be recognized as displaying, this information (Allwood 1976). In this paper, this kind of actions will be focused on. These actions are labeled *communicative motor actions* or simply *communicative actions*.

In the context of face-to-face communication, two basic types of communicative actions can be separated, *verbal actions,* i.e. speech actions alone, and *non-verbal actions*, e.g. facial or hand-arm actions that accompany or complement verbal actions. Other examples are co-verbal eye and head movements; however, these are not focused on in this paper. *Speech actions* result in a complex flow of temporally overlapping speech articulator movements (movements of lips, tongue, velum, and lower jaw) as well as laryngeal and sublaryngeal articulations which can be described as vocal tract action units (Saltzman and Munhall 1989; Goldstein et al. 2006, 2007). It will be argued in this paper that communicative actions usually bear a number of causally related goals on different levels of abstraction (e.g. from wanting to have a window closed, to signaling reference to a particular handle, to performing a pointing gesture). As with every communicative motor action, one immediate goal of speech actions is to produce an understandable signal, i.e. to transfer information between interlocutors. Such goals thus fall into the sensory domain where the goals of speech actions can be analyzed mainly in the *auditory domain* ("Communicative actions"), while those of non-verbal actions fall into the *visual domain*. In addition, the *somatosensory domain* plays a role for goal specification from the viewpoint of the speaker (Nasir and Ostry 2006), while the auditory and visual domains are important for the goal specification for the speaker as well as the recipient.

Speakers are often not aware of their non-verbal *facial expression*. Yet it is well known that *co-verbal facial actions* have a strong communicative function, for example by signaling the speaker's affective or emotional state (e.g. neutral, happy, sad, and angry), by signaling the speaker's connotation of the verbal message (e.g. seriousness vs. ironic), or by underlining important prosodic parts of an utterance (e.g. underlining the most stressed syllable of a sentence). Thus, a speaker's co-verbal facial actions which are perceived via the visual domain carry important information for the interlocutor in addition to the verbal message (Kopp et al. 2008). In the same line, speakers are often unaware of their co-verbal *gesturing*, i.e. of their *co-verbal hand-arm actions*. Similar to co-verbal facial expression, co-verbal gesturing can signal the affective or emotional state of the speaker. In addition, gesturing can supplement or complement the verbal massage of an utterance by, e.g. signaling a specific direction not expressed verbally through an additional hand-arm gesture. This latter type of co-verbal gesturing is the focus of the current paper.

The next section we summarize the basic principles and features of production, perception, and acquisition of willed motor actions. Then we illustrate that these basic principles and features apply to all types of communicative actions, i.e. to verbal and co-verbal communicative actions such as speech, facial, and hand-arm actions. Next we propose a model for the production, perception, and acquisition of communicative verbal and co-verbal actions will be proposed, which will act as a guide for structuring the computational implementation of production (or control) modules as well as perception-comprehension modules for embodied conversational agents, i.e. for humanoid robots as well as for three-dimensional virtual computer screen characters.

Basic principles in action production, action perception, and action acquisition

The literature identifies a set of basic principles or features for the production, perception, and acquisition of many types of (private and communicative) willed actions. It is argued in this paper that the main goal of communicative actions is *shape forming*. This applies to the whole or parts of the vocal tract in order to reach specific auditory goals for speech actions, or the face in order to reach specific visual goals for co-verbal facial actions, as well as for one or both hand-arm systems in order to reach specific visual goals for co-verbal hand-arm actions. Shape forming leads to specific *spatial or spatiotemporal targets* for all types of these actions. In this section, literature focusing on communicative gestures and private reaching and grasping actions is reviewed, while literature on private actions such as standing or walking is not included. Reaching and grasping actions serve as typical examples for willed actions in many studies on motor control (e.g. Arbib et al. 2000; Sabes 2000; Desmurget and Grafton 2000; Wolpert and Flanagan 2001; Nowak et al. 2007). In addition, literature on co-verbal gesturing (e.g. Kendon 2004), on speech actions ("vocal tract gestures", e.g. Saltzman and Munhall 1989; Browman and Goldstein 1989, 1992; Goldstein et al. 2006, 2007), and on facial actions ("facial action units", e.g. Cohn et al. 2007) will be considered to illustrate that the following basic principles are common for the production, perception, and acquisition of communicative actions in general.

- Action realization is hierarchically structured by (1) planning, i.e. specifying the target or goal of an action and by specifying features of the target-directed movement in distal space and time (spatial and temporal features), (2) articulator movement programming in proximal body coordinate space, and (3) action execution (see "Hierarchy of action representation and motor hierarchy").

- Actions can be separated with respect to function and behavior. Action function is represented by a formulation of the abstract and discrete action goal or action task. Behavior is represented by the quantitative articulator movements occurring during action execution and in the case of communicative actions by the resulting quantitative auditory and visual signals (section "Dichotomy of function and behavior of actions").
- Action programming and action execution comprise an abundance of redundant or equivalent task solutions. Performed task solutions are found by principles of synergy, optimality, or economy (section "Motor redundancy, motor abundance, motor equivalence, and motor synergy").
- Actions are learned or trained during action acquisition in the form of babbling and imitation training and can be adapted by further learning during the whole lifespan. Learned or trained actions (i.e. skilled actions) are executed mainly by feedforward control. Feedback control is needed primarily during learning and adaptation (section "Motor learning, feedforward and feedback control, internal models, and adaptation").
- Actions can be performed overtly (i.e. normal execution) or covertly (i.e. imagination of the action without movement generation). Covert as well as overt action production leads to activation of the neural representation of the (abstract) action goal and facilitates action understanding and reasoning (section "Overt and covert action performance, mirror system hypothesis, and action understanding").
- Action perception, i.e. the identification of the action goal by processing the articulator movement of the action, is a discrete process, and an action is often already understood before the target is reached. Thus, in many cases, an action target does not need to be reached explicitly during action execution (section "Static and dynamic information in action perception").

Hierarchy of action representation and motor hierarchy

Any action is organized hierarchically. *Action planning* starts with activation of the *abstract and discrete cognitive representation of the action's goal or task.* In the case of reaching a target, this is the activation of the symbolic representation, notion, or cognitive concept of "reaching" and of the object to be reached (Jeannerod 1999; Gallese 2000). This discrete planning level allows a separation of different actions and a discrete and distinctive description of each action (Lestou et al. 2008). Action planning continues with *quantitative spatiotemporal planning* (Sabes and Jordan 1997; Desmurget and Grafton 2000; Kawato 2000; Todorov 2004; Nowak et al. 2007). The distance between the position of the target-reaching articulator and the action target itself, the point in time for target reaching and thus the time-span intended for action execution, and eventually the optimal movement trajectory (i.e. a motion between the actual end-effector position and the target minimizing a certain cost function) as well as an optimal temporal profile for the movement (velocity profile) is estimated at this stage. Thus, a first rough spatiotemporal plan for the end-effector movement is planned in the so-called task space coordinate system (Saltzman 1979; Kelso et al. 1986), at least for some important spatiotemporal landmarks (Turvey 1977; Jordan 1995; Sabes 2000; Sober and Sabes 2005).

In the next step, this task space end-effector movement solution is broken down into movements for *all effectors*, i.e. all articulators, involved in the realization of an action (e.g. upper body, upper and lower arm, hand, thumb, and fingers in the case of reaching or hand gesturing; or upper lips, lower lips, and lower jaw in the case of a labial closing speech gesture). These effector movements are *programmed* on this level by trying to approximate the end-effector movement or by trying to approximate at least some spatiotemporal landmarks of this movement already specified during planning. Here, in addition, all knowledge concerning the muscular-skeletal subsystem for the actual effector is taken into account (e.g. how the effectors are coupled by joints and controlled by muscles or bundles of muscles and which neuromuscular activation leads to which effector-specific local movement). This *central or global motor program* of the action comprises a temporal coordination of all neuromuscular control signals of all effectors involved in the realization of an action. On the one hand, the motor program must be closely related to task space action movement planning. On the other hand, motor programming is effector-specific and muscle-oriented and takes into account all constraints concerning the muscular-skeletal system of the actor. On the effector level, movement descriptions are given in specific local joint-related coordinate systems (Saltzman 1979; Jordan 1995; Sober and Sabes 2003; Sober and Sabes 2005). The central motor program comprises a set of *local motor programs* or *motor commands* controlling all effectors involved in execution of an action (e.g. Kopp and Wachsmuth 2004). Since a motor command or local motor program can still be complex, these programs or commands can be broken down into a set of simpler motor commands realizing basic target-directed movement elements called *movement* *primitives* or *motor primitives* (Todorov and Ghahramani 2003).

The final level in the hierarchy of action representation is action *execution*. The motor commands or primitives for all effectors coordinated by the central or global motor program become activated and lead to a temporally coordinated neuromuscular activation and to coordinated movements of all effectors realizing an action (Saltzman and Munhall 1989; Jeannerod 1999; Sober and Sabes 2005). This level of action representation can be monitored, i.e. the resulting movements can be perceived by the actor himself and by others. From the viewpoint of the actor, the resulting movements can be monitored from visual, auditory, and somatosensory feedback. From the viewpoint of an observer or communication partner, executed action movements can be perceived in the visual or auditory domain.

Motor redundancy, motor abundance, motor equivalence, and motor synergy

The human muscular-skeletal apparatus has more degrees of freedom for executing a specific motor action than are necessary to master almost any given task or action. For example, in the case of reaching or grasping, the movements of the upper and lower arm and of the hand and fingers (i.e. *effectors* or *articulators)* must be coordinated. Depending on the action or task, an *end-effector* can be defined which has to perform the task (e.g. fingers in the case of grasping or pointing, upper and lower lips in the case of a bilabial closing speech action) while the other effectors (upper arm, lower arm, hand in the case of grasping, lower jaw in the case of many speech actions) are just needed to help perform the desired end-effector movement. Since the muscular-skeletal system normally comprises more degrees of freedom for potential effector movements than are necessary for mastering the task, a variety of seemingly equivalent movement alternatives could be generated, but one solution must be picked out during action planning. This occurrence of alternatives in task performance is called *motor redundancy* or *motor abundance* (Latash et al. 2008). A typical example for abundance is *motor equivalence*, i.e. to achieve a task in different ways if different specific perturbations or other changes or constraints occur for the body model (Abbs 1979; Kelso et al. 1984; Flash and Hogan 1985; Scholz et al. 2007). Redundancy can also occur on the neuromuscular level, since an effector movement can be performed by using different muscular activation strategies and since in the case of joints with many degrees of freedoms (e.g. shoulder, elbow joint, wrist in the case of hand-arm actions, mandibular joint, upper and lower lips movement degrees of freedoms in the case of bilabial speech actions) an effector movement can be realized by using different systems of muscles acting on the effectors. In addition, different movement contributions of different effectors can contribute to one single end-effector movement solution, e.g. jaw upper and lower lips can contribute in different portions to a bilabial closing action (Lindblom 1983).

The redundancy problem is solved on the motor programming level for example by *motor synergies* (Bernstein 1967), i.e. by the fact that effectors or articulators can work together in a coordinated way that maximizes efficiency of the resulting movement, while at the same time reduces the numbers of redundant degrees of freedoms and thus eases the control problem. In the case of redundancy for motor planning, *optimality control methods* (Todorov 2004) indicate how these redundancies can be decreased, i.e. how an optimal solution for performing the task can be found. A cost function for describing the overall cost of action performance can be defined in order to maximize *economy*, i.e. to minimize for example the overall muscle energy needed for performing an action (Rasmussen et al. 2001). In addition, it has been discussed whether it is important to minimize *smoothness* costs such as jerk or joint torque change in the case of simple target-directed actions (Nelson 1983; Hogan 1984; Todorov and Jordan 1998; Kawato et al. 1990; Smeets and Brenner 1999).

Motor learning, feedforward and feedback control, internal models, and adaptation:

Willed motor actions need to be trained or learned. Thus, these actions also can be labeled as *skilled movements* or *skilled actions* (Nelson 1983; Saltzman and Kelso 1987). Willed motor actions are improved until an optimal behavior is reached during action acquisition. Motor learning leads to experienced execution of actions as a result of trial-and-error motor productions. It can be hypothesized that an acting toddler starts building up a basic repertoire of *motor primitives* during *motor babbling* (Demiris and Dearden 2005; Der and Martinus 2006). The resulting motor primitives, here called *primitive actions,* are perceived by the toddler's feedback pathways in the visual, somatosensory, and in the case of vocal tract movements also in the auditory domain. Thus, the toddler is capable of acquiring basic *internal (forward) models* predicting the sensory results and the resulting movements for primitive actions from motor activation before action execution (Kawato 1999) by means of basic Hebbian learning. However, assuming a simple muscular-skeletal system with at least 30° of freedom and three motor primitives (forward, backward, and zero) per degree of freedom, one already arrives at an impracticable amount of 10^{14} potential primitive actions (Schaal 1999; Wolpert et al. 2001). This problem can be solved by incorporating *action imitation* at a very early stage in motor acquisition. Since imitated actions are composed from primitive actions, only those primitive actions are trained (or "babbled") which are needed later on for building up communicative and private actions. That way the toddler starts with the learning of simple actions and then proceeds with training of more and more complex actions (Sabes 2000; Iacoboni 2005; Nowak et al. 2007). The build-up of internal models during motor learning is a synonym for becoming more and more experienced or skilled in performing a specific action (Wolpert and Flanagan 2001). It is important to note that sensory feedback information is extensively used for learning, i.e. for building up internal models (Kawato 1999; Arbib et al. 2000; Desmurget and Grafton 2000). While the main goal of motor babbling is to train sensor-to-motor relations (i.e. to build up internal models), imitation training in addition associates motor plans and motor programs of skilled actions with meaning. The toddler not just imitates the movement pattern of a communication partner (e.g. caregiver) but in social contexts additionally becomes aware of the communicative or private function of the action to be trained, and thus links meaning or function of the action (i.e. discrete formulation of its goal) with its behavior (i.e. motor plan, articulator movements, and sensory outcome).

After learning or training, motor planning and programming of skilled actions can be done partly based on internal models, which accept copies of the motor commands (also called *efference copy*) and which predict likely sensory consequences. This allows motor planning and programming to be done mainly *before* the action is executed and is necessary since sensory feedback especially in the visual or acoustic domain has a significant delay. This process is called *feedforward control, open loop control,* or *anticipatory control.* In contrast, *feedback control* or *closed loop control* is used during action acquisition on all levels of motor control and during action execution at least on lower somatosensory levels of motor control (*servo mechanisms* or *servo controller*, Todorov 2004, p. 910; Sober and Sabes 2003, 2005).

It should be noted that higher level feedback control is not exerted for real-time control of the execution of an action but for evaluating the overall result of an action, e.g. whether the goal of the action was reached or not. But this kind of higher level feedback control allows (slow) *adaptation* of an action with respect to internal or external disturbances (e.g. internal perturbations like fixation of a joint resulting from a dysfunction or external perturbations like an obstacle between end-effector and action target; see Cheng and Sabes 2006) during several repetitions of that action.

Overt and covert action performance, mirror system hypothesis, and action understanding

Covert states of actions can be activated by imagining an action (*motor imagery*), by intending an action that will be eventually executed in future, or by observing an action performed by other individuals "as if the observer would use the implicit strategy of putting himself 'in the shoes of the agent'" (Jeannerad 2001, p. 104). Covert activation of actions can lead to a context-sensitive activation of nearly the whole (hierarchical) neural system needed for action performance, i.e. activation of characteristic neural areas of action performance within the primary motor cortex, premotor cortex, supplementary motor areas, prefrontal cortex, basal ganglia, and cerebellum as well as activation of the peripheral neural system (ibid., p. 104ff). However, the peripheral neural activity does not result in muscle activation, since the degree of motor unit neural activation is too small in this case for initiating muscular activation. In addition, in this case, neural inhibition mechanisms may help to prevent movement execution (ibid., p. 106). Thus, the covert state of action activation comprises anticipatory action planning and programming while *overt states* involve execution. Although both states lead to activation of nearly the same central areas, the overt state involves a higher level of activation of neuromuscular units (motor units) and in addition no inhibitory mechanisms on the neuromuscular level.

The covert activation of action planning is closely related to the *mirror neuron system concept* (Kohler et al. 2002; Fadiga and Craighero 2004; Rizzolatti and Craighero 2004), according to which a subordinate neural network links action perception and action production. This action resonance network yields activation of motor areas even if an action is just observed (covert state activation, e.g. Cunnington et al. 2006; Brass et al. 2007, Grafton and Hamilton 2007). In addition, it is hypothesized that *action understanding* is facilitated by the mirror neuron system, since in the case of known and already trained actions, action observation directly leads to an activation of the covert action state, which includes the activation of the high-level cognitive action specification, representing the action goal or action task in an abstract way (see "Hierarchy of action representation and motor hierarchy").

Static and dynamic information in action perception

On the one hand, it is evident for reaching or grasping actions as well as for co-verbal hand-arm actions that *kinematic and/or dynamic movement information* is essential for action perception and action understanding. Moreover, humans are capable of extracting the intention or the goal behind an action from observing the corresponding articulator movements, especially before the action-related articulator movements are completed, i.e. even before the goal of the action is fulfilled completely. It has been shown that human perception is specialized for the detection of *motion* of biological forms (Blakemore and Decety 2001). Furthermore, it has been shown that Fitt's law on the relation between speed and accuracy of action-based movements is not just an important motor control principle for action production but that this principle also holds for imagination and perception of actions (Grosjean et al. 2007). Experiments using point-light displays have also shown that pure movement information of the face is sufficient for classification of emotional expressions (Bassili 1978), that movement information of the mouth region increases speech perception in noise (Rosenblum et al. 1996), and that movement information of gesturing is important for gesture identification (Poizner et al. 1981). Thus, it can be assumed that a detailed description and a high-quality synthesis of the goal-directed *movement* behavior of actions is very important for producing and perceiving actions. Furthermore, humans become directly aware of "artificial" (i.e. low quality synthesis of) movement patterns, since natural movement patterns are the basis for action learning (Jastorff et al. 2006). In contrast, "artificial" static face or body components (artificial body shapes) and even unknown artificial creations of skeletal structures can be adapted easily in action learning (ibid.).

On the other hand, for some types of actions, the action targets are clearly reached and the present (static) target information can be a perception feature. However, inferring from these static action targets to a presumed prior communicative intention is not clear-cut and generally can only be achieved by integrating (static) target-signaling actions as sub-actions in a broader context (see "Actions, sub-actions, and primitive actions"). This holds for co-verbal hand-arm actions, e.g. if the target is to signal the size of an object by gesturing a distance between both hands, or for speech actions, e.g. in the case of manner of articulation for plosives and fricatives (i.e. the closure or critical constriction is reached and hold for a specific temporal interval). Notably, in these cases, the movement information of the actions is also used during perception. For example, the interlocutor becomes aware of co-verbal hand-arm actions during the movement part of these actions, and the place of articulation for plosives and fricatives is coded by formant transitions, i.e. by the vocal tract articulator movements. Even in the case of co-verbal facial actions, an emotion can be detected by static target information (i.e. still images of faces), but it has been shown that movement information delivers important additional information for emotion detection, e.g. whether a smile is natural or polite (Schmidt et al. 2006).

Dichotomy of function and behavior of actions

From the discussion of action hierarchy ("Hierarchy of action representation and motor hierarchy" and "Overt and covert action performance, mirror system hypothesis, and action understanding"), it can be concluded that it is advantageous to separate at least two perceptive levels of representation for action understanding, i.e. *action function* (meaning and form) and *action behavior* (movement), while three production levels can be separated for action representation, i.e. action *planning*, *programming*, and *execution*. The action function can be specified by describing the goal or task of an action in a discrete and abstract way (e.g. "grasp a cup" or "convey a message"). In the case of action observation, this abstract action goal or action function can also be labeled as action *meaning* and is comparable with the abstract and discrete level of action planning from the actor's viewpoint. Moreover, the *form* of an action can be noted in a discrete way (e.g. pointing by using the planar hand or stretched index finger). Action behavior comprises the physically measurable movement of all articulators or effectors involved in action execution. The resulting movement behavior is the basis for action perception in the somatosensory and visual domain and in the case of verbal actions also in the acoustic domain. The importance of separating action function and action behavior is, for example, reflected in specifying multimodal behavior generation for embodied conversational agents (Kopp et al. 2006), where a function markup language (FML) describing intent of communicative actions without referring to physical behavior is distinguished from a behavior markup language (BML) describing desired physical realizations of actions. From the neuroscientific view, it can be resumed that "the concept of willed actions allows the establishment of links between cognitive psychological models of control of action and the field of motor control." (Jahanshahi and Frith 1998, p. 484). Thus, from the viewpoint of cognitive psychology, an action can be described by *verbally specifying the goal of the action*. From the viewpoint of motor control, an action can be described as a *control unit of movement* leading to behavior.

Actions, sub-actions, and primitive actions

Complex actions themselves can be subdivided into *sub-actions* or *primitive actions* (Schaal 1999, p. 237; Grafton and Hamilton 2007). For example, a grasping action can be subdivided into two sub-actions, i.e. reaching the object with the hand-arm system and then grasping the object (e.g. a cup of coffee) by the hand system (thumb and digits). These two sub-actions are performed not strictly in a sequential manner since they overlap in time, i.e. formation of the hand for grasping the cup occurs during the time course of the reaching action. The same holds for speech actions, co-verbal facial actions, and co-verbal hand-arm actions (see "Communicative actions"). The simplest sub-action is called *primitive action* in this paper. In order to label an action as primitive, its goal has to be so simple that the appropriate action comprises not more than a simple target-directed end-effector movement in the case of speech or facial expression, or a simple target-directed or shape-conserving action in the case of depictive gesturing. For example, in this way, an iconic gesture is conceived of as composed of repeated circular movements as primitive actions; (see e.g. Kopp and Wachsmuth 2004, p. 47ff).

The difference between actions and sub-actions or primitive actions is that *actions (or super-ordinate actions) carry meaning*. The intention or meaning can be described in abstract and discrete categories like "grasping", "reaching", or in the case of communicative actions in abstract and discrete categories such as "transfer a word by speech", "signal an emotion by facial expression", or "indicate a distinct direction by gesturing". Thus, in the case of communicative actions, the goal is to convey *meaningful messages*, i.e. an understandable verbal speech item (sound, syllable, word, phrase, and utterance), by using vocal tract articulators, an understandable co-verbal gesture (e.g. *deictic actions* like pointing a direction or to an object or *iconic actions* displaying the shape of an object) by using the hand-arm articulators, or an affective or emotional state by using facial articulators.

For communicative actions, it will be exemplified later (section "Communicative actions") that primitive actions do not convey meaning but *information on discrete and abstract features*. Since an action is composed of a set of temporally well-coordinated primitive actions, the meaning of the (super-ordinate) action results from the complete set of features and their specific combination achieved through the structuring of the primitive actions. In the case of speech, an action represents the production of a word, phrase, or utterance, while a sub-action represents a syllable and a primitive action, a vocal tract action unit (see "Speech actions"). Feature information is here the syllabic sound chain or a bundle of distinctive features like place and manner of articulation. The features representing a sound or syllable are represented by the set of speech primitive actions realizing that sound or syllable. In the case of facial actions, typical feature information is for example mouth-corner raising/lowering or inner/outer eye brow raising/lowering which is controlled by facial primitive actions (cf. the notion of action units in the *facial action coding system* (Ekman and Friesen 1978), while a meaning is for example an emotional state as conveyed by a complete facial expression (see "Co-verbal facial actions"). In the case of gesturing, feature information is for example conveyed by primitive actions like retaining a particular shape of the hand or the direction of two successive linear movements, while the meaning of a gesture is conveyed by the stroke phase as a whole (see "Co-verbal hand-arm actions").

Communicative actions

"From the motor chauvinist's point of view, the entire purpose of the human brain is to produce movement. Movement is the only way we have of interacting with the world. All communication, including speech, sign language, gestures and writing, is mediated via the motor system." (Wolpert et al. 2001, p. 487). From this viewpoint, communicative actions can be interpreted as being as essential as the well-researched private actions such as reaching or grasping. In order to develop a specific action-based model for production, perception, and acquisition in face-to-face communication, the function and behavior of communicative actions, i.e. co-verbal facial and hand-arm actions as well as speech actions, will be discussed in detail in this section.

Speech actions

A theory of speech actions or speech gestures was conceptualized as coordinative structures in a dynamical perspective on speech production (Kelso et al. 1984, 1986) and further developed toward a quantitative control concept for speech articulation (Browman and Goldstein 1989, 1992; Saltzman and Munhall 1989; Saltzman and Byrd 2000; Goldstein et al. 2006, 2007). Articulatory Phonology (Browman and Goldstein 1989, 1992) postulates that *articulatory gestures* or *vocal tract action units* (Goldstein et al. 2006, 2007) are the basic or atomic units of speech production as well as for producing phonological contrast. A vocal tract action unit is, for example, a bilabial, apical, or dorsal closing gesture resulting in the realization of a bilabial, apical, or dorsal plosive or nasal, a vocalic tract forming gesture as is necessary for the production of vowels, a glottal opening or closing gesture performed in the production of voiceless or voiced sounds or a velopharyngeal opening or closing gesture as seen in the production of a nasal or a non-nasal (oral) speech sounds. Furthermore, "word forms are organized 'molecules' composed of multiple articulatory gestures (the 'atomic' units)" (Goldstein et al. 2006, p. 224). In terms of the definitions given in "Dichotomy of function and behavior of actions", vocal tract action units can be interpreted as *speech primitive actions,* while the organization or coordination of vocal tract action units into syllables, words, phrases, or utterances is considered as meaning-carrying *speech actions*. In the model proposed in this paper, speech primitive actions can thus be interpreted as contrastive or discriminative units, while speech actions are meaning-carrying units.

The planning stage of a speech action (e.g. a word) is represented in *phonological graphs* or *coupling graphs* (i.e. at the level of abstract, symbolic, and distinctive action representation, e.g. Browman and Goldstein 1989; Goldstein et al. 2006, p. 220; Kröger 1993, p. 221) and in *vocal tract action scores* or *gestural scores* (at the level of representation of temporal coordination of vocal tract actions, Browman and Goldstein 1989; Goldstein et al. 2006, p. 220; Kröger and Birkholz 2007, p. 184). The final state of action planning, i.e. the specification of action goals like target vocal tract shapes and movement trajectories of end-effectors is closely connected with the state of programming of *movements for all articulators* (or effectors) involved in the realization of a speech action. This quantitative state of action planning and programming is for example modeled in the task dynamics approach by separating inter-gestural and inter-articulatory coordination (Saltzman 1979; Saltzman and Kelso 1987; Saltzman and Munhall 1989; Saltzman and Byrd 2000). A detailed approach for modeling *neuromuscular control* of speech movements is the biomechanical model of Perrier et al. (1996, 2003), Payan and Perrier (1997). This approach is based on the equilibrium point hypothesis model of Feldman (1986) and additionally comprises higher level speech planning modules (Perrier and Ma 2008). Other biomechanical models of speech production have been proposed by Ito et al. (2004) and by Dang and Honda (2004).

In contrast to co-verbal facial and co-verbal hand-arm actions, where the goals can be specified in a direct movement-related domain (e.g. movement of facial skin points such as mouth corners in the case of facial actions or movement and form of the hands in the case of hand-arm actions) and where action goals are perceived by others within the visual domain, the domain of speech action goals is discussed controversially. On the one hand, vocal tract actions can be specified by movement-related goals formulated as location and degree of constriction in the vocal tract (tract variables, Saltzman and Munhall 1989). On the other hand, it is evident from the viewpoint of communication that the goal of speech actions is the transfer of information to others, which is mainly done by using the acoustic-auditory domain. Thus, it is argued that the goal of vocal tract actions should be formulated in the acoustic-auditory domain (Perkell et al. 1997; Guenther et al. 1998) or that the goal is multimodal with the acoustic-auditory modality having the highest level of priority (Perrier 2005). The multimodality of speech action goals is also supported by Nasir and Ostry (2008) claiming that speakers have precise somatosensory expectations independent of auditory goals.

Speech primitive actions can be separated with respect to onset, target (or steady-state), and offset phase (Kröger et al. 1995). No or only very short hold portions can be found for speech movements in normal or fluent speech production. Particularly, in the case of consonantal closures or constrictions, biomechanical saturation effects (Perkell et al. 1997) occur with respect to the collision of two articulators (e.g. lower and upper lips or vocal folds) or of an articulator with the vocal tract walls (e.g. tongue with palate), which result in steady-state phases of articulation. However, the center of mass of the articulator can still be moving in such cases. Speech articulation is thus characterized mainly by target-directed movements of articulators rather than by reaching targets through temporal target or steady-state phases. Furthermore, the idea of target undershoot (Lindblom 1963) suggests that targets of vocal tract action units are rarely reached at least in the case of vowels. Thus, in the case of normal fluent speech, speech primitive actions of vowels can be characterized as target-directed movements (i.e. by the onset phases of these actions) and rarely reach a sound target fully in fluent speech. The communicative goal of a speech primitive action, which is the discrimination or identification of sound features, does not necessarily result exclusively from constant steady-state auditory features but also from the auditory correlates of the (target-directed) movement pattern for both, consonants (e.g. formant transitions for place of articulation, Cooper et al. 1952; Kurowski and Blumstein 1984) and vowels (Strange et al. 1983; Nearey and Assmann 1986; Neel 2004). This underlines the importance of the onset phase in speech primitive actions.

Current models of speech production also focus on *speech acquisition* (Bailly 1997; Guenther 2006; Guenther et al. 2006; Kröger et al. 2009a, b) and confirm the importance of babbling and imitation which can be termed *vocal babbling* and *vocal imitation* in the context of this paper. As was stated in "Basic principles in action production, action perception, and action acquisition", acquisition of motor actions is strongly depending on sensory feedback. For speech actions, the importance of auditory and somatosensory feedback during speech acquisition is claimed in all these models. Furthermore, the overall importance of a production–perception link is stressed in several theories of speech production and speech perception (e.g. Liberman and Mattingly 1985; Hickok and Poeppel 2007; Schwartz et al. 2007).

Finally, it should be mentioned that although the visual outcome of speech actions, i.e. lip and jaw speech movements and the partially visible movements of the tongue, is a helpful cue in speech perception (e.g. Summerfield 1987), it is the primary goal of speech actions to produce understandable *acoustic* speech signals. *Facial* speech movements (i.e. visible speech movements like movements of the lips and of the lower jaw) should thus be seen as a by-product of speech actions, i.e. they are used as an additional cue in face-to-face communication by the speech perception system (e.g. if the acoustic signal is produced in noisy environments, e.g. Girin et al. 2001) but are not intended to be the primary cue of speech action identification.

Co-verbal facial actions

An emotional or affective state is expressed mainly and most directly by a facial expression. Facial expressions can be decomposed into anatomically or muscular-based "facial action units", which can be interpreted as the "smallest visually discriminable facial movements". (Cohn et al. 2007). A system of facial action units (facial action coding system, FACS) comprising goal-directed actions in the upper and lower face regions (eye brow, eye region, or mouth region action units) has been developed by Ekman and Friesen (1976, 1978) and was refined over the years (see Cohn et al. 2007). The facial action coding system has become the leading approach for representing and quantifying facial expressions. This system is capable of coding each occurring facial expression (e.g. prototypical facial expressions representing basic emotional states) by combinations of specific facial action units (Tian et al. 2005; Cohn 2007; Pantic and Rothkrantz 2000). Facial action units often occur in a temporally overlapping way. The combination of facial actions is "additive" if the appearance of each action unit is independent (no spatial overlap) or "non-additive" (i.e. complex interaction between action units), if these action units share the same facial regions and thus modify each other (Cohn et al. 2007).

Moreover, it has been shown that facial dynamics are very important in the perception of facial expressions (Ambadar et al. 2005), and it is argued that the facial action coding system is powerful in describing facial dynamics (Cohn 2007; Tian et al. 2005; Cohn et al. 2007). The temporal range of a facial action unit can be separated in onset, peak, and offset phase (de la Torre et al. 2007), and in the case of different action units representing one facial expression, these temporal phases do not need to occur in absolute synchrony. The importance of the kinematics of onset of facial action units has been demonstrated by Ambadar et al. (2005) for detecting subtle facial expressions, or by Schmidt et al. (2003, 2006, 2009) for detecting spontaneous vs. deliberate facial expressions.

In terms of the definitions given in "Dichotomy of function and behavior of actions", visual action units as defined in FACS can be interpreted as *facial primitive actions* or *co-verbal facial primitive actions*. A visual action unit specifies a visual change, e.g. a local shift of facial skin (like mouth corners) or the local emergence and local shift of wrinkles based on activations of a specific muscle or of a specific group of synergistically working muscles. The goals or targets of these anatomically based specific visible facial movements (e.g. raise/lower inner/outer eye brows, raise upper lid, tighten lid, raise cheeks, raise chin, wrinkle the nose, pull/depress lip corners, see Cohn et al. 2007) can be interpreted as facial visible features or appearance features (Tian et al. 2005), while meaning-carrying facial expressions (e.g. expressing emotional states or facial gestures) are *facial actions* or *co-verbal facial actions* in the model proposed in this paper. Thus, a facial expression or facial gesture results from a temporally coordinated combination of facial primitive actions (in most cases, strongly overlapping in time).

The facial musculature is fully formed and fully functional at birth. Newborns already show a considerable facial mobility (Ekman and Oster 1979, p. 533). Thus, it can be assumed that newborns are capable to perform *facial motor babbling*. Despite the fact that some facial expressions present in early infancy already resemble certain adult facial expressions, it has been found that 2- to 3-week-old infants can also imitate some facial movements (Ekman and Oster 1979, p. 534; Field et al. 1984; Meltzoff and Moore 1977, 1989). As children get older, they use their perceptual abilities to fine-tune their facial expressions (Schmidt and Cohn 2002, p. 12). Moreover, "preschool children know what the most common facial expressions look like, what they mean, and what kind of situations typically elicit them" (Ekman and Oster 1979, p. 534). Preschool children's spontaneous facial expressions also reflect the emotions shown by others (ibid., p. 535). It can thus be assumed that newborn and older children are capable of performing *facial action imitation,* resulting in learning the behavior and the function of facial actions.

Co-verbal hand-arm actions

Similar to co-verbal facial expressions, *co-verbal hand-arm actions*, also called *visible bodily actions* or *gestures* (Kendon 2004), convey information concerning the speaker's emotional and affective state. In addition, co-verbal hand-arm actions occur as "a part of the process of discourse, (or) as a part of uttering something to another in an explicit manner" (Kendon 2004, p. 1). This communicative use of gestures can occur with different types of co-verbal hand-arm actions (cf. McNeill 1992): (1) a pointing action for referring to something (deictic gestures), (2) a complex depictive action for displaying features of the shape of an object or event (iconic gestures) or an abstract idea or concept (metaphoric gestures), (3) small beat-like actions (Alibali et al. 2001) for emphasizing an important point (e.g. an important word) in the flow of speech, and (4) actions for signaling a question, a plea, a doubt, or actions for proposing a hypothesis, denying something, or indicating agreement (Kendon 2004, p. 1). Reflecting the structure of verbal discourse, co-verbal gestures tend to group in a *gesture unit* whose time course can be divided into one or more *gesture phrases* and a final *recovery phase* (Kendon 2004, p. 111ff), also called *retraction phase* (McNeill 1992). Each *gesture phrase* can be subdivided into the *preparation phase* and nucleus phase. The *nucleus phase* in turn can be subdivided into the *stroke phase* and not mandatory but frequently occurring *pre-* or *post-stroke hold phases*. Each gesture phrase is semantically or pragmatically related to an intonation phrase in speech (Kopp and Wachsmuth 2004), also called "tone unit" of the speech flow (Kendon 2004, p. 111ff). Moreover, the stroke phase of a gesture is temporally coordinated with the "tonic center" i.e. the primary pitch accent syllable, of the verbal intonation phrase or tone unit. The duration of the post-stroke hold is timed with respect to the duration of this tonic center or with respect to the length of the remaining tone unit or intonation phrase. Thus, the stroke phase plus post-stroke hold phase is temporally coordinated with the so-called "speech affiliate" (Kopp and Wachsmuth 2004). If the stroke is built up by one or more target-directed actions, the post-stroke hold is realized as a constant hold of the gesture target (ibid., p. 47ff). If the stroke action is an oscillatory action (e.g. signaling a rotation, ibid., p. 47ff), the post-stroke portion often comprises additional oscillations representing the target of that action (for speech gesture coordination see also Levelt et al. 1985; Rochet-Capellan et al. 2008).

Since the stroke phase and the post-stroke hold phase are the meaning-carrying units in co-verbal gesturing (conveying a unit of meaning or "idea unit", McNeill 1992), each stroke phase, together with the post-stroke hold phase, can be interpreted as one *co-verbal hand-arm action* in the terminological framework developed in "Dichotomy of function and behavior of actions". The stroke phase of a gesture itself can comprise a complex succession of sub-actions or primitive actions (e.g. a rapid inward-outward movement performed twice in order to signal "throwing", see Kendon 2004, p. 113ff, or a succession of an outward-downward and an inward-downward movement for signaling the shape of a window, see Kendon 2004, p. 116ff). These actions can be interpreted as *hand-arm primitive actions* or *co-verbal hand-arm primitive actions*, each realizing specific features of the gesture stroke phase. Notably, those primitive hand-arm actions are thus not considered meaningful themselves (McNeill 1992), in the sense that they cannot act like morphemes, but they nevertheless obtain distinct meaning features in the context of a holistic gestural image (cf. Kopp et al. 2007).

The preparation, stroke, and recovery phases of a gesture phrase as defined by Kendon (2004), p. 111ff should not be confused with the onset, target, and offset phases of a primitive hand-arm action. Similar to speech actions and co-verbal facial actions, co-verbal hand-arm actions always comprise a *set of temporally well-coordinated primitive actions*. The primitive actions of the stroke phase of a gesture phrase mainly carry specific features of the meaning of the gesture phrase, e.g. features specifying the direction in the case of a deictic gesture or a shape in the case of an iconic gesture (Kopp et al. 2007). However, the preparation phase of a gesture phrase is also realized by a co-verbal hand-arm primitive action, the goal (or action target) of which is e.g. to direct the hands to a specific position in the space between actor and interlocutor for starting the semantically important stroke phase. Furthermore, the retraction phase of a gesture unit is also realized by a co-verbal hand-arm primitive action, the goal (or action target) of which is e.g. to direct the hands to a specific rest position (folded hands, hands side by side, hands on a desk, arms and hands down beside the body, etc.).

The hierarchical control regime outlined in "Basic principles in action production, action perception, and action acquisition" mainly from literature on grasping or reaching actions can be explicitly adopted for co-verbal hand-arm actions as realized in Kopp and Wachsmuth's (2004) computational approach. Planning hand-arm actions is done (1) by selecting distinct gesture phrases by specifying abstract gesture features (e.g. hand shape, palm orientation, extension finger orientation, and goal of hand movement) and (2) by specifying the motor primitives required for the characteristic spatiotemporal and kinematic properties of the movement trajectories for at least the stroke phase of each gesture phrase. This is done in coordination with speech planning, in order to be able to synchronize gesture stroke and the speech affiliate in "chunks" (ibid., p. 43). Motor programming is done by specifying a *central or global motor control program* which is capable of controlling *lower level local motor programs* (as described in "Basic principles in action production, action perception, and action acquisition" of this paper). The flexibility obtained by adopting such a hierarchical action-based approach to co-verbal gesture generation as described here allows embodied conversational agents, e.g. to freely decide upon the meaning or function of their communicative actions, in fine coordination with those of their verbal actions, and then to realize all required gestural actions straightforward (Bergmann and Kopp 2009).

It is known that children begin to gesture before talking (Rodrigo et al. 2004) and go through different stages of gesture acquisition, i.e. the repertoire of gestures starts to develop from around 9 month of age and continues over preschool years up to the adult use of gestures (Guidetti and Nicoladis 2008). It can be assumed that co-verbal hand-arm gesturing profits from *general motor babbling* as occurring for the emergence of control over the whole human limb system (Demiris and Dearden 2005; Der and Martinus 2006). Furthermore, although gestures are often considered as originating from more or less successful praxic actions responded to by an observer (e.g. caregiver), it can be assumed that *hand-arm gesture imitation training* is important in the case of learning co-verbal hand-arm gestures. It should also be emphasized that children develop a gestural communication system very early in order to be able to interact with their caregivers before they are able to use speech (Tomasello et al. 2007). Thus, gesturing in general allows communication between young children and caregivers and facilitates language acquisition such as the development of the word lexicon (Guidetti and Nicoladis 2008).

A comprehensive action-based production, perception, and acquisition model for face-to-face communication

The research evidence discussed in the previous sections allows the conclusion that the concept of action can be applied to the overt behavior occurring in each domain of face-to-face communication, if actions are interpreted in the sense discussed in "Communicative actions" for speech, co-verbal facial expression, and hand-arm gesturing. In this section, a model for production, perception, and acquisition of actions in face-to-face communication will be proposed. This model comprises the following basic features:

- Each domain of face-to-face communication, i.e. speech, co-verbal facial expression, and co-verbal hand-arm gesturing, can be described using an *action-based concept* comprising *production, perception, and acquisition*. The concept of action is biologically motivated and thus comprises cognitive and sensorimotor components (see "Basic principles in action production, action perception, and action acquisition").
- Speech actions, co-verbal facial actions, and co-verbal hand-arm actions (i.e. communicative actions) are on the one hand functional in communication, i.e. they are *meaningful (or message carrying) units,* and on the other hand behavioral, i.e. specific movement units. Which specific functions and behaviors, as well as mappings between them, one can find in the verbal (speech) and co-verbal (hand-arm and facial) actions of face-to-face communication is likely to vary with language and cultural background.
- Communicative meaning-carrying *actions* (i.e. speech, co-verbal facial, or co-verbal hand-arm actions) are *sets of temporally coordinated primitive actions*. Primitive actions are simple target-directed or oscillatory actions. Oscillatory actions only occur in the domain of hand-arm gesturing, and their target is an ongoing oscillatory movement. Primitive actions are organized in *action scores* on temporal parallel tiers. They overlap temporally if they occur on different tiers (e.g. two facial primitive actions producing a co-verbal facial expression exhibit strong temporal overlap, vocalic tract shaping, and consonantal closing primitive actions producing a CV-syllable exhibit moderate temporal overlap) or they happen sequentially if they occur on one tier (e.g. co-verbal hand-arm primitive actions that exhibit small or no temporal overlap).
- Primitive actions comprise an *onset phase* for approaching the action target by the end-effector, eventually followed by a *target-, hold-, apex-, or peak-phase*, and eventually followed by an *offset phase*, in which the end-effector moves back to a neural position. In many cases, the onset phase of an action is directly followed by the onset phase of the following primitive action and thus suppresses the target and offset phase of the preceding primitive action.
- The *dynamic (or kinematic) features of onset behavior* of a primitive action allow the estimation of the action target. Onset duration, onset maximum movement amplitude, and onset maximum velocity are important kinematic parameters for action production and action perception.
- The immediate goal of communicative actions is to produce and to convey specific meaning-carrying *shapes*. A shape of the whole or of specific regions of the vocal tract and its acoustic-auditory correlates, i.e. the sound target, carry (verbal) sound feature information and is controlled by a set of speech primitive actions. A facial shape and its visual correlates including the dynamics for reaching this shape specify a meaning-carrying (co-verbal) facial expression and are controlled by the set of facial primitive actions. The shape of the stroke of a hand-arm gesture unit results from the dynamic and static aspects of the stroke, and its visual correlates carry meaning-carrying (co-verbal) gesture information and are controlled by a set of hand-arm primitive actions.
- In our action-based model, the physiological and neurophysiological *structure* can be separated from cognitive and sensorimotor *knowledge* about motor actions (i.e. action function and mapping to sensorimotor action

behavior). The knowledge results from learning or training (acquisition procedures).

- *Communicative actions* are *learned* or *trained* during speech acquisition and during the therefore necessary social face-to-face interaction processes occurring between toddler and caregiver or later between adult and interlocutor. *Babbling phase* and *imitation phase* are substantially different but equally important phases for the acquisition of communicative actions. *Action adaptation processes* occur during speech acquisition as well as after the initial training or learning phases during the whole lifetime.
- Action production and action perception are closely linked. The concept of action underlines this close *production–perception link* since actions are defined units of production (acquired sets of temporally coordinated primitive actions) as well as meaning-carrying units of perception. The close link of production and perception is supported by current neural theories of action processing (mirror system hypothesis).

Structure of the model

The model for production, perception, and acquisition of actions in face-to-face communication is biologically based and thus comprises cognitive, sensorimotor, and sensory parts (Fig. 1). The model operates in three basic functional modes, i.e. production, perception, and acquisition. The structure of the model is explained below by describing the production, perception, and acquisition mode.

Production

If an (abstract) message including emotional information is intended to be communicated by an actor, actions are selected in each domain (i.e. speech domain, co-verbal hand-arm domain, and co-verbal facial domain) for implementing the message conveyance process. Action selection is realized by activating meaning states and the appropriate form states in the *action function knowledge repository* (see "action function" in Fig. 1). This repository can also be called *action lexicon* or *mental lexicon*. The knowledge stored in this repository (i.e. meanings, forms, and meaning-form relations) is learned during action acquisition (see "Acquisition of action function, action behavior, and motor behavior knowledge"). In the speech domain, word meaning and the appropriate phonological words forms are selected and activated on the level of the mental lexicon for the realization of the intended utterance (Levelt et al. 1999; Indefrey and Level 2004). In the domain of gesturing, there is not a single clear-cut mental lexicon for all types of gestures. However, hand-arm actions (or gesture phrases) *and* primitive hand-arm actions (carrying features) can be collected in a gesture lexicon or *gestuary* (De Ruiter 1998, Kopp and Wachsmuth 2004) and can be selected, adapted, and combined depending on context (Kopp et al. 2007; Bergmann and Kopp 2009). Action selection (i.e. action activation) is done in all three domains by a cognitive process which activates neurons representing action meanings directly reflecting the context-dependent communicative intention

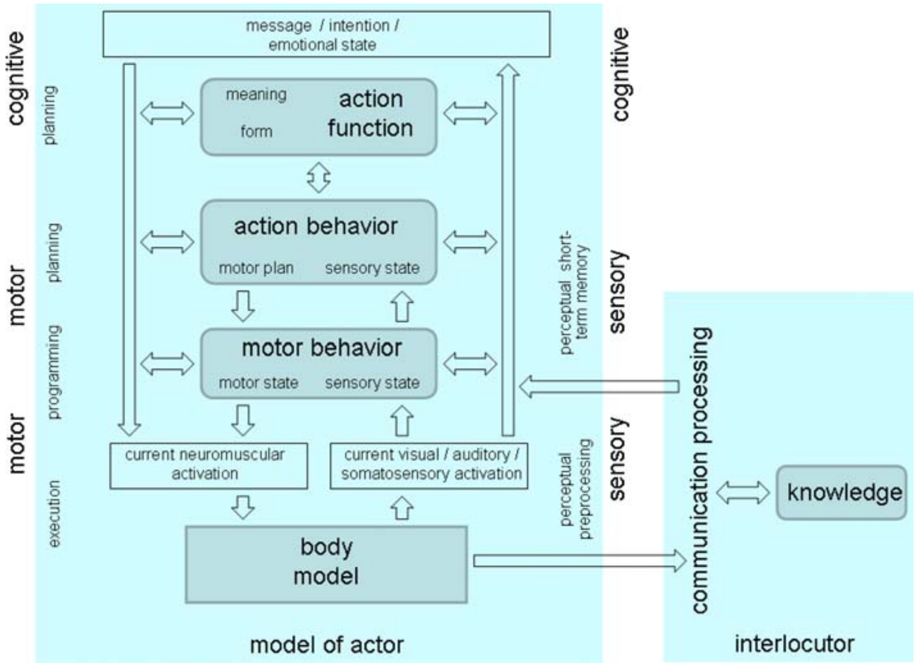

Fig. 1 Structure of the action-based model for face-to-face communication. *Shaded rounded boxes* represent knowledge repositories including processing of those cognitive, sensory, or motor states which are listed in these boxes. *Arrows* indicate information pathways which may include information processing

of the actor. Since an action can be seen as a unit, activation of neurons representing action meaning directly lead to co-activation of action form. For example, an action with intention or meaning "display the size of a small object" may lead to the activation of the form of a one-hand-two-finger gesture vs. a two-hand gesture for "display the size of a large object". In the domain of facial expressions, the intention or meaning "moderate happiness" may lead to the activation of the form or expression "little smile" vs. "big smile". Thus, the cognitive planning level consists in a selection and first coarse temporal ordering of discrete actions by "meaning of action" in each domain and leads to the activation of "form of the action", i.e. leads to an abstract and discrete description of the action goals and to a first coarse-grained temporal plan of the temporally co-occurring actions.

On the level of the *action behavior knowledge repository* (see "action behavior" in Fig. 1), the planning of the action leads to more and more quantitative descriptions by further specifying the temporal coordination of sub-actions and primitive actions realizing each domain-specific action, and then by specifying the spatiotemporal movement trajectory of the end-effectors for the whole set of (temporally coordinated) primitive actions. This specification is termed *motor plan* (*motor plan state* or *motor pattern*) in our model. In the domain of speech word, actions are built from one or more syllables (i.e. sub-actions). Motor plans (i.e. vocal tract action scores) display the temporal coordination of all speech primitive actions (i.e. vocal tract actions). For each primitive action, temporal parameters such as onset duration as well as spatial parameters such as a specific local or global vocal tract shape are additionally specified at this stage. In the domain of gesturing, the primitive actions are selected for all portions of all gesture phases (e.g. preparation, stroke, hold, and retraction), and the spatial and temporal constraints are specified firstly for the primitive actions of the stroke phase, and dependently follow for the primitive actions forming the other phases (cf. Kopp and Wachsmuth 2004). In the domain of facial expression, facial primitive actions (i.e. facial action units), their spatial target, and some temporal constraints (e.g. onset duration) are specified. It should be clear that actions in different domains (i.e. speech, facial expression, and gesturing) temporally co-occur. Furthermore, actions may overlap or co-occur even in one domain. This has already been illustrated for the speech domain. In the domain of gesturing, a co-occurrence of actions may occur if, for example, an iconic hand shape gesture is overlaid by a sentence stress underlining beat gesture.

Before the spatiotemporal movement trajectories of action-specific end-effectors are specified, the actions are temporally coordinated across domains at this stage (e.g. the speech affiliate and the stroke phase of the co-verbal gesture). The activation of spatiotemporal movement trajectories for end-effectors leads to co-activation of corresponding sensory states (somatosensory, auditory, and visual states) on the level of the action behavior knowledge repository (Fig. 1). The actor has activated planning so far at this stage that he/she is able to imagine internally how the execution is performed (activation of motor plan state), how this execution "feels like" (co-activation of appropriate somatosensory state), and how the execution will "sound" or "look like" (co-activation of appropriate auditory and visual state). These sensorimotor correlations are learned during action acquisition and labeled *action behavior knowledge* in our model (for an implementation of a multidirectional link of motor, sensory, and function states in speech see Kröger et al. 2009a).

Based on action hierarchy, it can be assumed that programming of effector movements, i.e. the specification of movement of all articulators in local joint coordinates and thus in terms of neuromuscular activation of specific muscles or bundles of muscles, is achieved on lower motor levels, subsumed as *motor behavior knowledge repository* in our model (see "motor behavior" in Fig. 1). The motor plan leads to a specification of a (central or) global motor program, which itself leads to a specification of lower level local motor programs and thus to temporally coordinated motor commands for all effectors involved in the realization of an action (Kopp and Wachsmuth 2004 for co-verbal hand-arm actions, Saltzman and Munhall 1989 for speech actions). Local motor behavior is trained during motor babbling and enforced during execution of any gesture. Thus, local sensorimotor knowledge exists for each local joint control regime. This knowledge enables the actor to predict and assess the sensory result of local effector movements controlled by local motor programs or motor commands already before movement execution. These lower level sensory estimates constrain the higher level motor planning and may lead to modifications of motor planning before action execution.

Facial primitive actions are directly controlled by one single muscle or one single group of synergetically working muscles. Only one local motor program is needed for the execution of a facial primitive action, directly controlling activity of a specific muscle or group of muscles. But it should be noted that the production of a facial action representing a specific emotional expression in many cases comprises the temporal coordination of more than one facial primitive action.

Perception

Internal somatosensory and auditory signals of the actor are used as feedback signals (*self-perception*). These signals result from action execution via the body model (Fig. 1).

Somatosensory feedback allows for a partly local control of production on different levels of programming and planning. On the motor programming or motor behavior level, somatosensory signals help to control local effector movements. On the action behavior level, the somatosensory, auditory, and visual signals help to control the execution of the end-effector movement in order to reach the goal of an action, i.e. to ensure that the interlocutor is able to understand the intended message. On this level of planning, the feedback signals also help to control the correct temporal coordination of actions of different domains (e.g. to control the temporal coordination of gesturing and speech). Internal feedback control further helps to detect incorrect productions of an action. If severe productions errors occur, i.e. if the communicative goal may not to be reached, the system can interrupt the flow of utterances and can start high-level self-correction procedures (e.g. to repeat an utterance) as a result of sensory feedback.

External visual and auditory signals produced by communication partners (interlocutors) are the basis for perception, understanding, and coordination in face-to-face communication. For action understanding, these external signals generated by a communication partner are compared on the side of the actor (now: recipient) with sensory states (or patterns) of already learned actions (stored actions). With respect to the assumption of a close production–perception link, this comparison leads to activation of action candidates during perception. The most activated stored action becomes the "perceived" action. This may result in a co-activation of the appropriate motor state, which then leads to a co-activation of the appropriate functional state (i.e. form and meaning state) for this action (this approach is exemplified for V- and CV-syllable speech actions by Kröger et al. 2009a, for hand-arm gesture actions by Sadeghipour and Kopp 2009). Thus, sensory signals produced by communication partners are preprocessed and forwarded to the short-term memory of the perceiver (the actor in Fig. 1) for activating sensory states or sensory patterns of his/her pre-learned actions. The activated and thus selected actions on the behavioral level lead to automatic and implicit segmentation and discretization of the incoming continuous flow of visual and auditory information. On the cognitive level, the activation of stored actions described earlier leads to an activation of meaning candidates and thus action understanding. This model hence also elucidates the influence of action learning or action training during action acquisition on perception.

Acquisition of action function, action behavior, and motor behavior knowledge

In addition to the structure of the model, the detail of the knowledge incorporated on all levels of the model (i.e. shadowed rounded boxes in Fig. 1) is crucial for the quality reached by computer-implemented versions of this model. This knowledge is gained during acquisition procedures. Two basically different phases or modes occur in the acquisition of communicative actions, babbling phase and imitation phase. During *motor babbling training,* the motor system of the actor (i.e. of the toddler) produces random movements and is capable of mapping the sensory consequences of these movements with the appropriate motor states. This results in collecting motor behavior knowledge. After babbling, the model is capable of predicting motor states from sensory states, and vice versa. Since it is not economic to learn the sensorimotor relations for all movements, which can potentially be generated by trying all combinations of all local effector movement primitives, it can be assumed that the toddler starts with *imitation training* very early on. Importantly, this starts with the caregiver imitating the motor babbling of the toddler, which provides the toddler with correlations between own diverse motor actions and sensory input patterns about other's actions. By generalizing from these correlations, the toddler now can try to imitate actions produced by communication partners or caregivers (interlocutor in Fig. 1). The external sensory input pattern given by the action produced by the communication partner can now directly be used for choosing a related motor pattern.

After execution of one motor pattern candidate, the actor has to evaluate whether the result of imitation (i.e. the current sensory pattern generated by the actor) is acceptable or not. This evaluation can be done in part by self-assessment (e.g. by comparing sensory cues of the self-produced and the external action) but is done mainly socially by assessing the current reaction of the communication partner on the produced imitation stimulus. Thus, the communication partner has to perform communication processing (Fig. 1). This early start of imitation training directs babbling into "typical" or "effective" body movement primitives for communication as are needed for producing meaning- or message-carrying actions. Thus, the set of training items for babbling and imitation training is shaped by the knowledge of communication partners (Fig. 1) during speech acquisition. This knowledge can be interpreted in technical terms as a stimulus database for training the model. The result of action acquisition is the build-up of knowledge concerning motor behavior, action behavior, and the functioning of actions.

Adaptation results from further learning of the model even if the basic acquisition phases mentioned earlier are completed. This further learning leads to modifications on the level of action behavior and motor behavior (Fig. 1). If, for example, the status of the body model changes as a result of a specific dysfunction (e.g. fixed lower jaw, fixed hand-arm joints, or facial hemiplegia) or an external perturbation (e.g. speaking while eating, external shift of formants, or gesturing with an object in one hand), motor plans and motor programs have to be changed in order to reach a comparable sensory result for action understanding. If, on the one hand, the change of status of the body model just affects non-end-effectors (e.g. jaw or upper arm) motor programming is mainly affected. Since a certain degree of flexibility in effector movement realization during action execution is already trained during action acquisition, and hence already stored as knowledge on the motor behavior level (Fig. 1), adaptive solutions may occur nearly in real time and without much additional training (i.e. without the need for many trials). For example, in the case of a lower jaw fixed by a bite block, adaptation occurs immediately and without training (Fowler and Turvey 1981). If, on the other hand, an external perturbation affects the movement of end-effectors or their resulting visual or auditory output, an adaptation process takes place over several trials for each affected action and thus needs time. In this case, the comparison of current and stored sensory states for a specific action on the action behavior level (Fig. 1) leads to a noticeable mismatch, which initiates adaptation processes by modifying the motor plan to sensory state relations on the action behavior level. Adaptation is complete if the relation of motor plan and sensory state is shifted far enough for the current sensory pattern of action performance to match with the expected stored sensory pattern for this action. In the same way, an "after-effect" occurs if an external perturbation, which the actor already has adapted to, is switched off. This is the case, for example, if a constant formant shift is applied to the perceptual system of the actor, affecting the formant trajectories which reflect end-effector movement trajectories in the acoustic domain. Here, adaptation requires learning time and in addition, a noticeable after-effect occurs (Houde and Jordan 2002; Purcell and Munhall 2006).

Discussion and conclusion

It was the goal of this paper to demonstrate that the *concept of action* commonly employed as a sensorimotor control concept for transitive or private actions such as grasping or reaching (e.g. Jahanshahi and Frith 1998; Jeanerod 1999; Todorov 2004; Latash 2008) is also useful for the sensorimotor description of those willed actions that make up face-to-face communication. With an application to the domain of speech production, of co-verbal facial expression production, and of co-verbal hand-arm gesture production, an action-based model for face-to-face communication has been proposed, which underlines the close connection between production, perception, and acquisition of actions. This model is also biologically and neurophysiologically based.

Two different types of actions are postulated in this paper: *Meaning-carrying actions* are accomplished by specifying a motor plan which comprises a temporally coordinated ensemble of *primitive actions*. For facial primitive actions, it can be assumed that these facial action units are units of production as well as of perception, since these actions are also characterized as "smallest visibly discriminable units" (Cohn et al. 2007). In the case of speech, there is evidence for the importance of acoustic sound features in speech perception (e.g. Diehl et al. 2004). Sound features can be interpreted as basic functional discrete information for specifying speech primitive actions in our approach. Due to the mirror system hypothesis, models exist which postulate the co-activation of motor states during speech perception (cf. dorsal pathway, Hickok and Poeppel 2007). In addition, there is evidence that meaning-carrying portions of the acoustic signal also can be processed as a whole (cf. ventral pathway, ibid.). Thus, it can be assumed that both units, i.e. meaning-carrying actions as well as primitive actions, play a certain role in the complex process of speech perception, of facial expression perception, as well as of hand-arm gesture perception (see horizontal arrows on the perception side of the actor model in Fig. 1). From a more theoretical viewpoint, it should be noted that an action-based approach can afford a hypermodal representation of sensory, motor, and functional states in a straightforward way (see the relation between action function and action behavior in Fig. 1). It has been shown for speech that such a hypermodal representation (called phonetic map in the case of speech, see Kröger et al. 2009a) closely connecting function, motor, and sensory states allows to explain basic effects of speech perception such as strong categorical perception for consonants and weak categorical perception for vowels.

It has been shown that the action goal is in the auditory domain in the case of speech and in the visual domain in the case of co-verbal facial expressions and co-verbal hand-arm gestures. With respect to the postulated close relation of production, perception, and functional states of actions, the notion of motor goals is transferable also to communicative actions. The goal of communicative actions is always to be sufficiently successful in transferring information to communication partners (or interlocutors). This goal is defined on the function side of any action but implies a sensory pattern and a motor pattern that is learned or trained during action acquisition. Thus, motor goals and sensory goals (i.e. auditory, visual, and somatosensory goals) are closely related on the action behavior level (Fig. 1). Also, the notion of action goals proposed here provides a basis for grounding the higher semantic or pragmatic structures of communicative intent planning that eventually control the behavior of interlocutors. If the function state of an action is activated on the cognitive level, a co-activation of the appropriate motor and sensory states occurs (for speech see Kröger et al. 2009a). Thus, motor goals and sensory goals are closely related, and the question whether an action goal is more on the motor or sensory side becomes an ill-posed question in the context of this approach.

Finally, action-based concepts are promising for computational models of the recognition as well as the synthesis of speech, facial expressions, and gesturing (quantitative computational approaches which incorporate a lot of ideas outlined in this paper are given by Steels and Spranger 2008 or by Kopp, to appear). Facial expression recognition using dynamic visual features and using a recognition strategy based on facial action units has great potential if high-quality data are available (Tian et al. 2005). A flexible motor-based approach is advantageous if high-quality synthesis of co-verbal gesturing is the goal (e.g. Kopp and Wachsmuth 2004): it is not possible to synthesize natural looking hand-arm gestures from a simple rigid movement database, since gesture realizations have to be flexible, for example, with respect to length and intonation variations of the verbal phrase produced in parallel. An integrated, biologically and neurophysiologically based concept of communicative actions as outlined in this paper can serve as a guide for designing recognition and synthesis applications in face-to-face communication.

Acknowledgments This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG) Project Nr. Kr 1439/13-1 and project Nr. Kr 1439/15-1, and by the Deutsche Forschungsgemeinschaft (DFG) in SFB 673 "Alignment in Communication" and the Center of Excellence "Cognitive Interaction Technology" (CITEC).

References

Abbs JH (1979) Speech motor equivalence: the need for a multi-level control model. In: Proceedings of the ninth international congress of phonetic sciences, Institute of Phonetics, Copenhagen, pp 318–324

Alibali MW, Heat DC, Myers HJ (2001) Effects of visibility between speaker and listener on gesture production. J Memory Lang 44:169–188

Allwood J (1976) Linguistic communication as action and cooperation. Gothenburg monographs in linguistics 2. Göteborg University, Department of Linguistics, Göteborg

Ambadar Z, Schooler J, Cohn JF (2005) Deciphering the enigmatic face: the importance of facial dynamics to interpreting subtle facial expressions. Psychol Sci 16:403–410

Arbib MA, Billard A, Iacoboni M, Oztop E (2000) Synthetic brain imaging: grasping, mirror neurons and imitation. Neural Netw 13:975–997

Bailly G (1997) Learning to speak: sensory-motor control of speech movements. Speech Commun 22:251–267

Bassili JN (1978) Facial motion in the perception of faces and of emotional expression. J Exp Psychol Hum Percept Perform 4:373–379

Bergmann K, Kopp S (2009) Increasing the expressiveness of virtual agents—autonomous generation of speech and gesture for spatial description tasks. In: Proceedings of 8th international conference on autonomous agents and multiagent systems (AAMAS 2009), pp 361–368

Bernstein N (1967) The coordination and regulation of movement. Pergamon, London

Blakemore SJ, Decety J (2001) From the perception of action to the understanding of intention. Nat Rev Neurosci 2:561–567

Brass M, Schmitt RM, Spengler S, Gergely G (2007) Investigating action understanding: inferential processes versus action simulation. Curr Biol 17:2117–2121

Browman C, Goldstein L (1989) Articulatory gestures as phonological units. Phonology 6:201–251

Browman C, Goldstein L (1992) Articulatory phonology: an overview. Phonetica 49:155–180

Cheng S, Sabes PN (2006) Modeling sensorimotor learning with linear dynamical systems. Neural Comput 18:760–793

Cohn JF (2007) Foundations of human computing: facial expression and emotion. In: Huang TS, Nijholt A, Pantic M, Pentland A (eds) Artificial intelligence for human computing (LNAI 4451. Springer, Berlin, pp 1–16

Cohn JF, Ambadar Z, Ekman P (2007) Observer-based measurement of facial expression with the facial action coding system. In: Coan JA, Allen JJB (eds) Handbook of emotion elicitation and assessment. Oxford University Press, New York, pp 203–221

Cooper F, Delattre P, Liberman A, Borst J, Gerstman L (1952) Some experiments on the perception of synthetic speech sounds. J Acoust Soc Am 24:597–606

Cunnington R, Windischberger C, Robinson S, Moser E (2006) The selection of intended actions and the observation of others' actions: a time-resolved fMRI study. NeuroImage 29:1294–1302

Dang J, Honda K (2004) Construction and control of a physiological articulatory model. J Acoust Soc Am 115:853–870

De la Torre F, Campoy J, Ambadar Z, Cohn JF (2007) Temporal segmentation of facial behavior. In: Proceedings of the IEEE 11th international conference on computer vision (ICCV 2007), Rio de Janeiro, Brazil, pp 1–8

De Ruiter JP (1998) Gesture and gesture production. Doctoral dissertation at Catholic University of Nijmegen, The Netherlands (now called Radboud University Nijmegen)

Demiris Y, Dearden A (2005) From motor babbling to hierarchical learning by imitation: a robot developmental pathway. In: Berthouze L, Kaplan F, Kozima H, Yano H, Konczak J, Metta G, Nadel J, Sandini G, Stojanov G, Balkenius C (eds) Proceedings of the fifth international workshop on epigenetic robotics: modeling cognitive development in robotic systems, Lund University Cognitive Studies 123, Lund, Sweden, pp 31–37

Der R, Martinus G (2006) From motor babbling to purposive actions: emerging self-exploration in a dynamical systems approach to early robot development. In: S Nolfi, G Baldassarre, R Calabretta, JCT Hallam, D Marocco, JA Meyer, O Miglino, D Parisi (eds) From animals to Animats 9. Proceedings of the 9th international conference on simulation of adaptive behavior (SAB 2006, Rome, Italy) LNAI 4905, Springer, Heidelberg, pp 406–421

Desmurget M, Grafton ST (2000) Forward modeling allows feedback control for fast reaching movements. Trends Cogn Sci 4:423–431

Diehl RL, Lotto AJ, Holt LL (2004) Speech perception. Annu Rev Psychol 55:149–179

Ekman P, Friesen WV (1976) Measuring facial movement. Env Psychol Nonverbal Behav 1:56–75

Ekman P, Friesen WV (1978) Facial action coding system. Consulting Psychologists Press, Palo Alto

Ekman P, Oster H (1979) Facial expressions of emotion. Annu Rev Psychol 30:527–554

Fadiga L, Craighero L (2004) Electrophysiology of action representation. J Clin Neurophysiol 21:157–168

Feldman AG (1986) Once more on equilibrium point hypothesis for motor control. J Mot Behav 18:17–54

Field TM, Woodson R, Greenberg R, Cohen D (1984) Discrimination and imitation of facial expressions by neonates. In: Chess S, Thomas A (eds) Annual progress in child psychiatry and child development. Brunner, Mazel, New York

Flash T, Hogan KN (1985) The coordinate of arm movements: an experimentally confirmed mathematical model. J Neurosci 5:1688–1703

Fowler CA, Turvey MT (1981) Immediate compensation in bite-block speech. Phonetica 37:306–326

Gallese V (2000) The inner sense of action: agency and motor representations. J Conscious Stud 7:23–40

Girin L, Schwartz JL, Feng G (2001) Audio-visual enhancement of speech in noise. J Acoust Soc Am 109:3007–3020

Goldstein L, Byrd D, Saltzman E (2006) The role of vocal tract action units in understanding the evolution of phonology. In: Arbib MA (ed) Action to language via the mirror neuron system. Cambridge University Press, Cambridge, pp 215–249

Goldstein L, Pouplier M, Chen L, Saltzman L, Byrd D (2007) Dynamic action units slip in speech production errors. Cognition 103:386–412

Grafton ST, Hamilton AF (2007) Evidence for a distributed hierarchy of action representation in the brain. Hum Mov Sci 26:590–616

Grosjean M, Shiffrar M, Knoblich G (2007) Fitts' law holds for action perception. Psychol Sci 18:95–99

Guenther FH (2006) Cortical interaction underlying the production of speech sounds. J Commun Disord 39:350–365

Guenther FH, Hampson M, Johnson D (1998) A theoretical investigation of reference frames for the planning of speech movements. Psychol Rev 105:611–633

Guenther FH, Ghosh SS, Tourville JA (2006) Neural modeling and imaging of the cortical interactions underlying syllable production. Brain Lang 96:280–301

Guidetti M, Nicoladis E (2008) Introduction to special issue: gestures and communicative development. First Language 28:107–115

Hickok G, Poeppel D (2007) Towards a functional neuroanatomy of speech perception. Trends Cogn Sci 4:131–138

Hogan N (1984) An organizing principle for a class of voluntary movements. J Neurosci 4:2745–2754

Houde JF, Jordan MI (2002) Sensorimotor adaptation of speech I: compensation and adaptation. J Speech Lang Hear Res 45:295–310

Iacoboni M (2005) Neural mechanisms of imitation. Curr Opin Neurobiol 15:632–637

Indefrey W, Level PJM (2004) The spatial and temporal signatures of word production components. Cognition 92:101–144

Ito T, Gomi H, Honda M (2004) Dynamical simulation of speech cooperative articulation by muscle linkages. Biol Cybern 91:275–282

Jahanshahi M, Frith CD (1998) Willed action and its impairments. Cogn Neuropsychol 15:483–533

Jastorff J, Kourtzi Z, Giese MA (2006) Learning to discriminate complex movements: biological versus artificial trajectories. J Vis 6:791–804

Jeannerad M (2001) Neural simulation of action: a unifying mechanism for motor cognition. NeuroImage 14:S103–S109

Jeannerod M (1999) The 25th Bartlett lecture: to act or not to act: perspectives on the representation of actions. Q J Exp Psychol 52A:1–29

Jordan MI (1995) Computational aspects of motor control and motor learning. In: Heuer H, Prinz W, Keele SW, Bridgeman B (eds) Handbook of perception and action: motor skills. Academic Press, London, pp 71–120

Kawato M (1999) Internal models for motor control and trajectory planning. Curr Opin Neurobiol 9:718–727

Kawato M, Maeda Y, Uno Y, Suzuki R (1990) Trajectory formation of arm movement by cascade neural network model based on minimum torque-change criterion. Biol Cybern 62:275–288

Kelso JAS, Tuller BT, Vatikiotis-Baetson E, Fowler CA (1984) Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures. J Exp Psychol Hum Percept Perform 10:812–832

Kelso JAS, Saltzman E, Tuller B (1986) The dynamical perspective on speech production: data and theory. J Phon 14:29–59

Kendon A (2004) Gesture: visible action as utterance. Cambridge University Press, New York

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

Kopp S (to appear) Social resonance and embodied coordination in face-to-face conversational with artificial interlocutors, speech communication (special issue on speech and face-to-face communication)

Kopp S, Wachsmuth I (2004) Synthesizing multimodal utterances for conversational agents. J Comput Anim Virtual Worlds 15:39–51

Kopp S, Krenn B, Marsella S, Marshall AN, Pelachaud C, Pirker H, Thórisson KR, Vilhjálmsson H (2006) Towards a common framework for multimodal generation: the behavior markup language. In: Gratch J, Young M, Aylett R, Ballin D, Olivier P (eds) Intelligent virtual agents (LNCS 4133. Springer, Berlin, pp 205–217

Kopp S, Tepper P, Ferriman K, Cassell J (2007) Trading spaces—how humans and humanoids use speech and gesture to give directions. In: Nishida T (ed) Conversational informatics. Wiley, Oxford, pp 133–160

Kopp S, Allwood J, Ahlsen E, Grammer K, Stocksmeier T (2008) Modeling embodied feedback in a virtual human. In: Wachsmuth I, Knoblich G (eds) Modeling communication with robots and virtual humans (LNAI 4930. Springer, Berlin, pp 18–37

Kröger BJ (1993) A gestural production model and its application to reduction in German. Phonetica 50:213–233

Kröger BJ, Birkholz P (2007) A gesture-based concept for speech movement control in articulatory speech synthesis. In: Esposito A, Faundez-Zanuy M, Keller E, Marinaro M (eds) Verbal and nonverbal communication behaviours, LNAI 4775. Springer, Berlin, pp 174–189

Kröger BJ, Schröder G, Opgen-Rhein C (1995) A gesture-based dynamic model describing articulatory movement data. J Acoust Soc Am 98:1878–1889

Kröger BJ, Kannampuzha J, Neuschaefer-Rube C (2009a) Towards a neurocomputational model of speech production and perception. Speech Commun 51:793–809

Kröger BJ, Kannampuzha J, Lowit A, Neuschaefer-Rube C (2009b) Phonetotopy within a neurocomputational model of speech production and speech acquisition. In: Fuchs S, Loevenbruck H, Pape D, Perrier P (eds) Some aspects of speech and the brain. Peter Lang, Frankfurt, pp 59–90

Kurowski K, Blumstein SE (1984) Perceptual integration of the murmur and formant transitions for place of articulation in nasal consonants. J Acoust Soc Am 73:383–390

Latash ML (2008) Evolution of motor control: from reflexes and motor programs to the equilibrium-point hypothesis. J Hum Kinet 19:3–24

Latash ML, Gorniak S, Zatsiorsky VM (2008) Hierarchies of synergies in human movements. Kinesiology 40:29–38

Lestou V, Pollick FE, Kourtzi Z (2008) Neural substrates for action understanding at different description levels in the human brain. J Cogn Neurosci 20:324–341

Levelt WJM, Richardson G, Heij WL (1985) Pointing and voicing in deictic expressions. J Memory Lang 24:133–164

Levelt WJM, Roelofs A, Meyer AS (1999) A theory of lexical access in speech production. Behav Brain Sci 22:1–38

Liberman AM, Mattingly IG (1985) The motor theory of speech perception revised. Cognition 21:1–36

Lindblom B (1963) Spectrographic study of vowel reduction. J Acoust Soc Am 35:1773–1779

Lindblom B (1983) Economy of speech gestures. In: McNeilage PF (ed) The production of speech. Springer, New York, pp 217–245

McNeill D (1992) Hand and mind: what gestures reveal about thought. University of Chicago Press, Chicago

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

Meltzoff AN, Moore MK (1989) Imitation in newborn infants: exploring the range of gestures imitted and the underlying mechanisms. Dev Psychol 25:954–962

Nasir SM, Ostry DJ (2006) Somatosensory precision in speech production. Curr Biol 16:1918–1923

Nasir SM, Ostry DJ (2008) Speech motor learning in profoundly deaf adults. Nat Neurosci 11:1217–1222

Nearey T, Assmann P (1986) Modeling the role of inherent spectral change in vowel identification. J Acoust Soc Am 80:1297–1308

Neel AT (2004) Formant detail needed for vowel identification. Acoust Res Lett Online 5:125–131

Nelson WL (1983) Physical principles for economics of skilled movements. Biol Cybern 46:135–147

Nowak DA, Topka H, Timmann D, Boecker H, Hermsdörfer J (2007) The role of the cerebellum or predictive control of grasping. Cerebellum 6:7–17

Pantic M, Rothkrantz LJM (2000) Expert system for automatic analysis of facial expressions. Image Vis Comput 18:881–905

Payan Y, Perrier P (1997) Synthesis of V-V sequences with a 2D biomechanical tongue model controlled by the equilibrium point hypothesis. Speech Commun 22:185–205

Perkell J, Matthies M, Lane H, Guenther F, Wilhelms-Tricarico R, Wozniak J, Guiod P (1997) Speech motor control: acoustic goals, saturation effects, auditory feedback and internal models. Speech Commun 22:227–249

Perrier P (2005) Control and representation in speech production. ZAS Pap Linguist 40:109–132

Perrier P, Ma L (2008) Speech planning for VCV sequences: influence of the planned sequence. In: Proceedings of the 8th international seminar on speech production, Strasbourg, France, pp 69–72

Perrier P, Ostry DJ, Laboissiere R (1996) The equilibrium point hypothesis and its application to speech motor control. J Speech Hear Res 39:365–378

Perrier P, Payan Y, Zandipour M, Perkell J (2003) Influence of tongue biomechanics on speech movements during the production of velar stop consonants: a modeling study. J Acoust Soc Am 114:1582–1599

Poizner H, Bellugi U, Lutes-Driscoll V (1981) Perception of American sign language in dynamic point-light displays. J Exp Psychol Hum Percept Perform 7:430–440

Purcell DW, Munhall KG (2006) Adaptive control of vowel formant frequency: evidence from real-time formant manipulation. J Acoust Soc Am 120:966–977

Rasmussen J, Damsgaard M, Voigt M (2001) Muscle recruitment by the min/max criterion—a comparative numerical study. J Biomech 34:409–415

Rizzolatti G, Craighero L (2004) The mirror neuron system. Annu Rev Neurosci 27:169–192

Rochet-Capellan A, Laboissiere R, Galvan A, Schwartz JL (2008) The speech focus position effect on jaw-finger coordination in a pointing task. J Speech Lang Hear Res 51:1507–1521

Rodrigo MJ, Gonzalez A, de Vega M, Muneton-Ayala M, Rodriguez G (2004) From gestural to verbal deixis: a longitudinal study with Spanish infants and toddlers. First Lang 24:71–90

Rosenblum LD, Johnson JA, Saldana HM (1996) Point-light displays enhance comprehension of speech in noise. J Speech Hear Res 39:1159–1170

Sabes PN (2000) The planning and control of reaching movements. Curr Opin Neurobiol 10:740–746

Sabes PN, Jordan MI (1997) Obstacle avoidance and a perturbation sensitivity model for motor planning. J Neurosci 17:7119–7128

Sadeghipour A, Kopp S (2009) A probabilistic model of motor resonance for embodied gesture perception. In: Proceedings of intelligent virtual agents (IVA09), pp 80–103

Saltzman E (1979) Levels of sensorimotor representation. J Math Psychol 20:91–163

Saltzman E, Byrd D (2000) Task-dynamics of gestural timing: phase windows and multifrequency rhythms. Hum Mov Sci 19:499–526

Saltzman E, Kelso JAS (1987) Skilled actions: a task dynamic approach. Psychol Rev 94:84–106

Saltzman E, Munhall KG (1989) A dynamic approach to gestural patterning in speech production. Ecol Psychol 1:333–382

Schaal S (1999) Is imitation learning the route to humanoid robots? Trends Cogn Sci 3:233–242

Schmidt KL, Cohn JF (2002) Human facial expressions as adaptations: evolutionary questions in facial expression research. Am J Phys Anthropol 116(S33):3–24

Schmidt KL, Cohn JF, Tian Y (2003) Signal characteristics of spontaneous facial expressions: automatic movement in solitary and social smiles. Biol Psychol 65:49–66

Schmidt KL, Ambadar Z, Cohn JF, Reed LI (2006) Movement differences between deliberate and spontaneous facial expressions: zygomaticus major action in smiling. J Nonverbal Behav 30:37–52

Schmidt KL, Bhattacharya S, Denlinger R (2009) Comparison of deliberate and spontaneous facial movement in smiles and eyebrow raises. J Nonverbal Behav 33:35–45

Scholz JP, SChöner G, Hsu WL, Jeka JJ, Horak F, Martin V (2007) Motor equivalent control of the center of mass in response to support surface perturbations. Exp Brain Res 80:163–179

Schwartz JL, Boe LJ, Abry C (2007) Linking dispersion-focalization theory and the maximum utilization of the available distance features principle in a perception-for-action-control theory. In: Sole MJ (ed) Experimental approaches to phonology. Oxford University Press, Oxford

Shadmehr R, Mussa-Ivaldi FA (1994) Adaptive representation of dynamics during learning of a motor task. J Neurosci 14:3208–3224

Smeets JB, Brenner EA (1999) A new view on grasping. Mot Control 3:237–271

Sober SJ, Sabes PN (2003) Multisensory integration during motor planning. J Neurosci 23:6982–6992

Sober SJ, Sabes PN (2005) Flexible strategies for sensory integration during motor planning. Nat Neurosci 8:490–497

Steels L, Spranger M (2008) The robot in the mirror. Connect Sci 20:337–358

Strange W, Jenkins J, Johnson T (1983) Dynamic specification of coarticulated vowels. J Acoust Soc Am 74:695–705

Summerfield Q (1987) Some preliminaries to a comprehensive account of audio-visual speech perception. In: Dodd B, Campbell R (eds) Hearing by eye: the psychology of lipreading. Lawrence Erlbaum, London, pp 3–51

Tian YL, Kanade T, Cohn JF (2005) Facial expression analysis. In: Li SZ, Jain AK (eds) Handbook of face recognition. Springer, New York, pp 247–275

Todorov E (2004) Optimality principles in sensorimotor control. Nat Neurosci 7:907–915

Todorov E, Ghahramani Z (2003) Unsupervised learning of sensory-motor primitives. In: Proceedings of the 25th annual international conference of the IEEE engineering in medicine and biology society, pp 1750–1753

Todorov E, Jordan MI (1998) Smoothness maximization along a predefined path accurately predicts the speed profiles of complex arm movements. J Neurophysiol 80:696–714

Tomasello M, Carpenter M, Liszkowski U (2007) A new look at infant pointing. Child Dev 78:705–722

Turvey MT (1977) Preliminaries to a theory of action with reference to vision. In: Shaw R, Bransford J (eds) Perceiving, acting and knowing: towards an ecological psychology. Erlbaum, Hillsdale, pp 211–266

Wolpert DM, Flanagan JR (2001) Motor prediction. Curr Biol 11:R729–R732

Wolpert DM, Ghahramani Z, Flanagan JR (2001) Perspectives and problems in motor learning. Trends Cogn Sci 5:487–494