Review article · JBR · 1-12Proofreading version

Towards an Articulation-Based Developmental Robotics Approach for Word Processing in Face-to-Face Communication

Bernd J. Kröger, [1]* Peter Birkholz,[1] Christiane Neuschaefer-Rube [1]

1 Department of Phoniatrics, Pedaudiology, and Communication Disorders, RWTH Aachen University, Aachen, GERMANY

Received ...

Accepted ...

Abstract

While we are capable of modeling the shape, e.g. face, arms, etc. of humanoid robots in a nearly natural or human-like way, it is much more difficult to generate human-like facial or body movements and human-like behavior like e.g. speaking and co-speech gesturing. In this paper it will be argued for a developmental robotics approach for learning to speak. On the basis of current literature a blueprint of a brain model will be outlined for this kind of robots and preliminary scenarios for knowledge acquisition will be described. Furthermore it will be illustrated that natural speech acquisition mainly results from learning during face-to-face communication and it will be argued that learning to speak should be based on human-robot face-to-face communication. Here the human acts like a caretaker or teacher and the robot acts like a speech-acquiring toddler. This is a fruitful basic scenario not only for learning to speak, but also for learning to communicate in general, including to produce co-verbal manual gestures and to produce co-verbal facial expressions.

Keywords

developmental robotics · humanoid robotics · conversational agents · face-to-face-communication · speech · speech acquisition · speech production · speech perception

1. Introduction

While humanoid face-to-face communication robots are currently under development in many labs and while the body structure of these robots is already very human-like – or at least human-like enough to be accepted and perceived as an artificial human being by human communication partners – the control principles of these robots are not. At present, rule-based artificial intelligence approaches are mainly used to control cognitive processes as well as sensory and motor processes in face-to-face communication systems. Rule-based approaches basically do not include learning processes. But humans acquire their knowledge for accomplishing communication processes – as well as other behavioral processes – on the entire amount of interactions with the environment, i.e. (i) on the entire set of environmental impressions including the actions of communication partners they perceived during their lifetime and (ii) on the entire set of all bodily actions and reactions (e.g. manual, facial, and speech actions) they produce during their lifetime (Tomasello 2000, Lungarella et al. 2003, Kuhl 2004, Kuhl 2007, Asada et al. 2009).

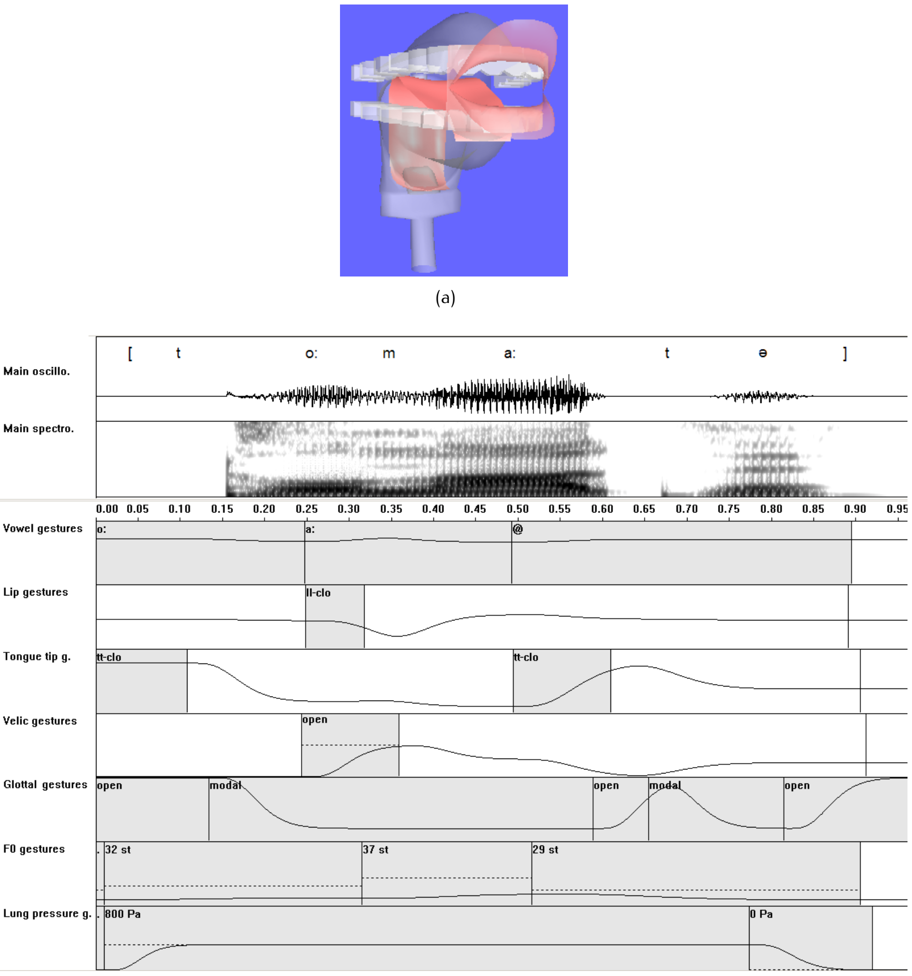
In this paper it will be argued that control module (i.e. the "brain model") and plant (i.e. the "body" including abstractions of arms, hands, specific parts of the face, and speech organs) should be divided in a way that the plant can directly be modeled with respect to a human archetype (i.e. *genetically based knowledge*), while the knowledge – which must be "uploaded" to the control module or brain model (i.e. *epigenetically based knowledge*) – has to be learned or acquired from a huge training set of human-robot interactions in a comparable way as humans themselves acquire their behavioral knowledge (cf. Weng et al. 2001, Prince and Demiris 2003, Weng 2004). In contrast to humans this complex process of knowledge acquisition needs to be done only for one robot exemplar and the acquired knowledge then can be simply "uploaded" to other robots, if they are intended to be used in comparable communication scenarios.

After discussing the importance of the facial, the manual, as well as the vocal tract domain in face-to-face communication (chapter 2) and after discussing basic principles for controlling a face-to-face interactive humanoid robot (chapter 3) the state of the art concerning humanoid communicative robots will be outlined (chapter 4). Thereafter, on the basis of current literature, a feasible basic architecture (i.e. a blueprint) for the control module of a humanoid robot specialized in face-to-face speech communication will be outlined (chapter 5) and subsequently a hypothetical basic training scenario will be described for word learning (chapter 6).

2. The domains of face-to-face communication

If we assume two persons which are communicating with each other face-to-face, the basic tasks are (i) to perceive and comprehend *communicative actions* produced by the other and (ii) to react on these actions, i.e. to produce adequate communicative actions for continuing the communication process with respect to the communicative goals (i.e. intentions) of each partner (e.g. Vilhjálmsson 2009). Communicative actions can be speech actions (i.e. verbal actions), as well as co-verbal facial expression actions, or co-verbal manual ges-

*E-mail: bkroeger@ukaachen.de



(a)

(b)

Figure 1. *(a) Software realization of a vocal tract plant comprising lips, tongue, velum, upper and lower jaw, pharyngeal wall, and larynx, for a speaker of Standard German (Birkholz and Kröger 2006) and (b) a control scheme for articulator movements (i.e. speech action score) realizing the German word "Tomate". From top: phonetic transcription, oscillogram, spectrogram, time scale in seconds, and articulator movement trajectories for tongue height, lip aperture, tongue tip height, velopharyngeal aperture, glottal aperture, vocal cord tension, and lung pressure. Activation intervals of vocal tract actions are marked as light gray boxes: three vocalic actions, one labial and two apical (tongue tip) closing actions, one velopharyngeal opening action, three glottal opening and two phonatory (modal) actions, and actions for adjusting fundamental frequency and lung pressure occur; articulatory movement trajectories are calculated using the dynamic model introduced by Birkholz et al. (in press).*

tures. Thus, three ***articulatory domains*** are important in face-to-face speech communication, i.e. the ***vocal tract domain*** comprising the oral, nasal, velopharyngeal, and laryngeal region with its articulators (e.g. lips, lower jaw, tongue, velum, glottis) in order to produce an acoustic speech signal, the ***facial domain*** comprising the eye region, cheeks, mouth and chin etc. for producing co-verbal facial expressions, and the ***manual domain*** comprising arms, hands, and fingers in order to produce co-verbal gesturing (Kröger and Kopp et al. 2010).

Moreover three ***perceptual domains*** can be differentiated, i.e. the auditory, the visual, and the somatosensory domain. While articulator movements of the vocal tract domain are mainly perceived in the ***auditory domain*** (since it is the goal of these movements to produce distinct acoustic signals), articulatory movements of the facial and manual domain (i.e. movements of the eye brows, eyelids, cheeks etc., or of the arms, hands, and fingers) are perceived in the ***visual domain***. Vocal tract actions at least of the lips, the lower jaw and the anterior part of the tongue can be perceived in the visual domain as well (e.g. "lip reading") and this kind of visual speech perception influences overall speech perception (see "McGurk-effect", McGurk and MacDonald 1976). But auditory perception is clearly dominant in the case of speech, since verbal communication can be performed successfully by exclusively using the auditory signal path without including the visual path (e.g. conversation by telephone) but not vice versa.

Conversational agents or robots mainly use the acoustic and auditory domain for modeling speech production and perception without introducing articulation, i.e. modeling of vocal tract articulator movements, while in the case of co-verbal manual and facial actions, production always implies modeling the *generation of movements* and perception always implies the visual *analysis of movements* or at least of spatial visible target configurations resulting from movements. It is a main idea of our approach to understand the acoustic speech signal as a signal which results from the *movement* of articulators (Fig. 1), in the same way as the visual signals occurring in the co-verbal facial and manual domain result from facial and manual articulator movements (see the unified theory for verbal and co-verbal communicative actions introduced by Kröger and Kopp et al. 2010). It will be shown in this paper that an *articulation-based interpretation* of speech production and perception – in parallel to the movement-based production and perception of manual and facial actions – is an essential and indispensable feature of any biological plausible model of speech communication.

Last but not least it is important to mention the *somatosensory domain* as the perceptual domain for monitoring the execution of actions produced by the robot or actor itself in each articulatory domain. This monitoring comprises tactile sensation (e.g. lips, hard palate in the case of speech articulation) as well as proprioceptive sensation; e.g. sensation of muscular tension for example in order to perceive the positioning of the tongue or sensation of degree of joint bending for example in the case of the lower jaw. On the one hand in the case of speech actions (i.e. vocal tract actions) it is well known that – beside auditory feedback – somatosensory feedback is important for controlling speech articulation (Golfinopoulos et al. 2011). On the other hand manual actions are controlled by somatosensory as well as by visual feedback (i.e. visual perception of the movements of the actors own hands and fingers) during their acquisition process (Iverson et al. 1999, Saunders and Knill 2004, Desmurget and Grafton 2000) while later on manual actions are mainly controlled by somatosensory feedback in face-to-face communication processes.

3. Self-organization and associative learning as basic principles

Associative and self-organizing neural network approaches are biologically plausible for controlling human behavior, but not yet implemented successfully in either humanoid robots or artificial agents involved in human-machine communication. Nevertheless, in this paper it will be argued that associative and self-organizing neural network approaches should be used, because these approaches are closely related to the biologically realistic functional processes occurring in the human brain (Thompson 1986, Kohonen 2001, Grossberg 2010) and thus potentially allow a high degree of naturalness in controlling communication processes.

A control module can be called an *associative control module*, if two conditions apply. (i) Stimulus exposure during learning is dual and synchronous. That is the case, if for example an auditory *and* a visual stimulus are exposed synchronously to the robot or toddler as is the case in specific word learning scenarios (Plebe et al. 2010, Goldstein et al. 2010), or if a motor pattern of an action and the sensory pattern, which results from the execution of that action, are exposed synchronously to the robot or toddler, as is the case in babbling training (Guenther et al. 2006, Kröger et al. 2009). (ii) An associative learning rule must govern the learning process, resulting in successful co-activation e.g. of an auditory (word) pattern if the visual pattern of an object is activated (Plebe et al. 2010) or e.g. of motor-patterns if an appropriate perceptual stimulus is activated (Kröger et al. 2009). Associative learning has been demonstrated to be a main biological principle for behavior learning (Mitchell et al. 2009) and is assumed as a basic principle especially in combined sub-symbolic and symbolic processing (Haikonen 2009).

A controller can be called *self-organizing control module*, if (i) there exist no predefined hardwired control rules, and (ii) if learning is unsupervised and learning results in adaptive behavior. A main feature of self-organizing control modules is that they reflect an ordering and categorization of behavior (e.g. speech, manual or facial actions) with respect to the main features which describe the variety of the behavior in each domain; e.g. phonetic features in the case of speech (Kröger et al. 2009) or movement primitives in the case of hand-arm actions (Tani et al. 2008, Tani and Ito 2003). A second feature of a self-organizing control module is that the representation of knowledge for a group of similar behaviors is larger the stronger the module is exposed to this group of stimuli during training. Both features of self-organization occur in human brains (Trappenberg et al. 2009, Grossberg 2010).

In communication processes as well as in many other behavioral processes it is important to subdivide cognitive and sensorimotor processing. *Cognitive processing* mainly acts on *symbolic items* (e.g. semantic concepts or phonological descriptions of words) while *sensory and motor processing* mainly acts on *sub-symbolic items* like motor or movement patterns or like visual, auditory, or somatosensory patterns. An associative and self-organizing control approach can be used in order to model sub-symbolic (i.e. sensory and motor) *as well as* symbolic (i.e. cognitive) processing; see Haikonen (2009) for a general discussion of symbolic and sub-symbolic processing and see Kröger and Kopp et al. (2010) for the unification of sub-symbolic and symbolic representations in communicative actions. In the next chapter, typical architectures of communicative agents or robots are described. All these architectures in principle can be implemented by using associative, adaptive, and self-organizing neural network approaches.

4. Autonomous communicative robots and their control: the state of the art

Face-to-face communication needs two autonomous subjects (e.g. an agent or robot and a human) capable of interacting with each other. This does not necessarily mean that these subjects have available a common language. For example two persons with different language backgrounds are capable of communicating and are capable of exchanging information more or less successful by nonverbal actions (e.g. facial expressions and manual gestures). Steels (2003) reports that two autonomous agents, each equipped with a cognitive system (i.e. a system processing symbolic information), with a sensory system (i.e. a system perceiving and processing sensory information, e.g. visual information concerning objects occurring within the robot's environment), and with a motor system (i.e. a system for performing actions by using the robot's effectors; e.g. head, arms, hands, fingers) are capable of developing a shared communication system. But the “evolving” language is not necessarily as complex as human languages are. Since the coded information can be communicated from robot to robot only by the effectors the robots have available, the kind of embodiment determines the “phonetics” of the evolving language: For example communication can be performed by eye- (or camera-) pointing to objects or by using specific gestures (see also Cangelosi and Riga 2006, Galantucci and Steels 2008).

Parisi (2010) suggests a human robot model comprising a linguistic and a non-linguistic neural sub-network, each composed of a sensory part, a motor part, and an intermediate layer for processing internal units (i.e. a cognitive part). In the case of the non-linguistic sub-network, the sensory part is capable of processing visual information (e.g. objects in the robot's environment) and the motor part is capable of accomplishing actions using its effectors (e.g. reaching and/or grasping an object by using specific effectors). In the case of the linguistic sub-network, the sensory part processes auditory information (e.g. speech items produced by the robot itself or by another robot or person in the robot's environment) and the motor part is capable of producing speech items by using vocal tract effectors (phono-articulatory organs). If the linguistic sub-network is activated the robot initially generates random movements of the vocal tract effectors, perceives the acoustic results of his own productions, and learns sensorimotor relations on the basis of these sensorimotor data (i.e. babbling; see also Kröger et al. 2009). In a second step the robot is capable of imitating speech items produced by another robot or by a "caretaker" or "robot-sitter" like a baby tries to imitate its caretaker s words and utterances. A comparable behavior result from activity of the non-linguistic sub-network: During a "motor babbling" period, the robot is capable of learning sensorimotor relations concerning the robot's body effector system as a basis for later perceiving or performing specific bodily actions (reaching, grasping etc; see also Demiris and Dearden 2005, Caligiore et al. 2008, Schaal 1999). Moreover Parisi (2010) discusses potential connections between the cognitive parts of both sub-networks and emphasizes that the occurring ***associations between the non-linguistic and linguistic cognitive sub-networks*** can be interpreted as a basis for linguistic comprehension (i.e. identifying the meaning of an utterance with respect to the linguistic and environmental context) as well as for non-linguistic comprehension (e.g. to comprehend an action or gestures or e.g. to notice the arrangement of an ensemble of objects).

Madden et al. (2009) describes a hybrid comprehension model comprising three core modules: a ***situated simulation module*** for internally simulating and representing action sequences up to shared plans for coordinated actions of two or more actors, a ***sensorimotor front-end module*** (perception/action module), and a ***cognitive module*** for processing non-linguistic as well as linguistic intentions, e.g. action planning/comprehension as well as utterance planning/comprehension (predicate-argument module). The important feature of this approach is the situated simulation module, which connects the cognitive symbolic and the sensorimotor modules within the model. This module can be activated from both sides, i.e. from the sensorimotor side for accomplishing the perception as well as the production of specific action sequences or from the cognitive part in order to concretize intentions (i.e. to produce actions or action sequences) or to comprehend external events (i.e. external situations as well as external speech).

While the robot control approaches introduced above more generally cope with ***production and comprehension of communicative actions and language***, the importance of face-to-face interaction as a basic vehicle for human communication in general (Grossmann et al. 2008) as well as for language performance and language learning in particular (Tomasello 2000, Dohen et al. 2010) guides us now to a discussion of front-end systems which can be called ***conversational*** or ***communicative robots or agents***, which are especially designed for ***face-to-face communication*** (e.g. Kopp et al. 2005, Bailly et al. 2010). These robots or agents can be seen as a sub-group of ***humanoid robots*** (examples for autonomous humanoid robots ***not*** specialized in face-to-face communication but dealing with human-robot interaction are given by Kanda et al. 2004 and Kanda et al. 2008, and by Kosuge and Hirata 2004). A main feature of face-to-face communicative robots or agents is their ability of mutual facial gazing including production and perception of facial expressions, of head gestures, and of eye movements (e.g. Rich et al. 2010, Sidner et al. 2005). Facial expressions, head, body and manual gestures, eye-movements, etc. can also be called backchanneling signals, if these signals are produced by the interlocutor (e.g. Ogawa and Watanabe 2000, Fujie et al. 2004). These speaker-listener signals are important for regulating the ongoing dialogue for example in order to signal the degree of engagement or cooperative behavior (Rich et al. 2010, Kanda et al. 2007), to regulate turn taking (Yoshikawa et al. 2006, Shiwa et al. 2008) and last but not least to monitor the current emotional state of speaker or listener (e.g. via differences in facial expressions, e.g. Hashimoto et al. 2010, Shiomi et al. 2004). At least ***sociable agents*** or ***sociable robots*** including cognitive and emotional control systems have been postulated and constructed in order to provide face-to-face communicative robots not just with cognitive but as well with social and emotional competence in order to make them appear as a socially and emotionally better understandable and predictable interlocutor in human-robot communication scenarios (Brooks et al. 1999, Breazeal 2003 and 2004, Bergman and Kopp 2009, Kopp et al. 2009).

The main problem for establishing a humanoid robot specialized in face-to-face communication is to provide the robot with typical human-like control knowledge. Thus the problem of establishing humanoid communication robots is tightly connected with solving the problem of modeling the ***autonomous development of the mental system***, i.e. solving the problem of developing behavior as well as of developing internal mental representations on the basis of ongoing lifelong learning (Weng et al. 2001, Prince and Demiris 2003, Weng 2004). It is widely accepted that the physical ***brain and body structure*** as well as a specific ***intrinsic developmental program*** is predefined (genetically defined). A main goal of developmental robotics is to stimulate ***lifelong learning*** from this intrinsic developmental program. The resulting (lifelong) training "events" should not be predefined in detail by this intrinsic developmental program but should result from this program as well as from the not necessarily full predictable interaction of the robot with its environment; at least the learning subject or robot should be capable of stimulating the occurrence of specific learning situations (Lindblom and Ziemke 2003, Asada et al. 2009).

Focusing on ***speech*** acquisition a major problem of current developmental robotics is that – even while the importance of sensorimotor interaction of the robot with its environment and even while the importance of embodiment is widely accepted – most robot architectures – even if they are used for research in developmental robotics of speech acquisition (e.g. Brandl 2009, Vaz et al. 2009) – just comprise an acoustically based but not an articulation based speech production and speech perception approach. First robotic vocal tract realizations are already existing (e.g. Fukui et al. 2005) but no attempts have been done to date in order to use these robots in the field of developmental robots for speech acquisition. Since the embodiment of the vocal tract apparatus is very important e.g. for human sensorimotor explorations occurring during speech acquisition (Kröger et al. 2009) as well as for natural modeling of speech production (Guenther et al. 2006, Golfinopoulos et al. 2011) and speech perception (e.g. Hickok and Poeppel 2007), it is the goal of this paper to develop a feasible brain model (chapter 5) and a hypothetical face-to-face communication training scenario (chapter 6) capable for modeling speech acquisition within the paradigm of developmental robotics.

5. A blueprint for a robot's speech processing "brain structure"

A brain model for speech communication should comprise lower-level processing routines for the ***articulation*** and for the ***perception*** of speech as well as some basic higher-level routines for the ***comprehen-***

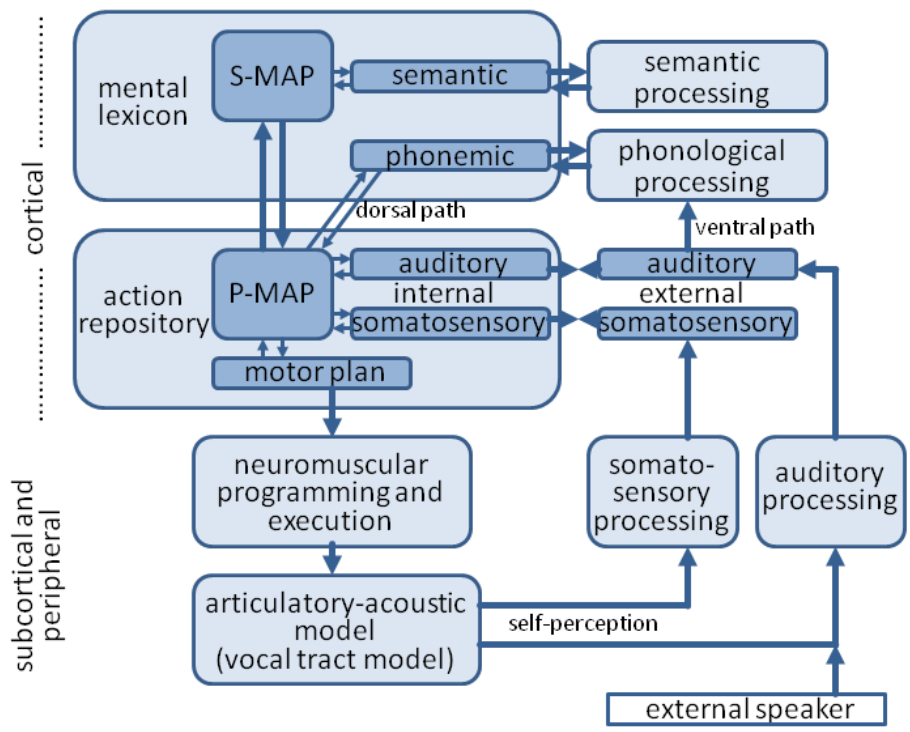


Figure 2. Blueprint of a brain model for speech production, speech perception, and speech acquisition. Light blue boxes indicate processing modules, dark blue boxes indicate self-organizing maps (S-Map and P-Map) or neural state maps (semantic, phonemic, auditory, somatosensory, motor plan state map); see text.

sion of a perceived utterance as well as for the ***production*** (i.e. conceptualization and formulation) of speech. A feasible interface between higher-level and lower-level processing is the ***phonological representation*** of a word or utterance. During a production task the activation of a phonological representation follows lexical item activation, i.e. lexical retrieval, lexical selection, and syntactic processing (Lau et al. 2008) and thus can be seen as the result of conceptualization and formulation (Levelt et al. 1999, Indefrey and Levelt 2004). During a comprehension task the activation of a phonological representation directly follows domain-specific processing of input information, mainly auditory information in the case of speech. In the case of auditory speech input these activation patterns can follow the ventral or dorsal route of speech perception (Hickok and Poeppel 2007). The description of a lemma level (i.e. a level for syntactic markers, Levelt et al. 1999) and syntactic processing is beyond the scope of this paper.

A blueprint of a brain model for speech processing following these ideas is given in Fig. 2. Here, processing of symbolic states (i.e. states which can be represented by symbols; e.g. phonological or semantic states) occurs near the language specific symbolic knowledge repository, i.e. the ***mental lexicon***. A crucial part of the mental lexicon is its ***central self-organizing map***, i.e. its ***semantic map*** (S-Map). This map is interconnected in a bidirectional way with a domain-specific state map, here with the semantic ***state map*** (note that the semantic state map and the semantic map are different neural maps, Fig. 2). In the case of a non-abstract object (e.g. a visible object like a dog) or a non-abstract action (e.g. a visible action like walking) the ***semantic state map*** is capable of representing symbolic information stemming from different domain- or mode-specific areas, i.e. from sensory areas like the visual areal, processing its visual form, color of its coat, like the somatosensory areal, processing the impressions concerning the tactile feedback during fingering its coat, like the olfactory areal, processing its smell, and like the auditory areal, processing its barking and yowling, as well as from motor areas, which together with visual areas process movement. The ***phonemic state map***, also appearing at the level of the mental lexicon (Fig. 2), is capable of representing language specific symbolic phonological information concerning the word; e.g. number of syllables, structure of each syllable (e.g. how many consonants occur in the onset and rhyme of the syllable), and phonological features of each sound within syllable onset and rhyme (e.g. manner and place of articulation). Following Li et al. (2004) both state maps occurring at the level of the mental lexicon, i.e. the semantic and the phonemic state map, are interconnected with two central self-organizing maps, named semantic map (S-Map) and phonetic map (P-Map). In addition to Li et al. (2004) the lower-level self-organizing map, i.e. the phonetic map, is also interconnected with sub-symbolic motor and sensory state maps and that this lower part is named action repository (Fig. 2). Consequently, this implies that phonemic states are closely related to sub-symbolic phonetic motor plan and sensory (i.e. auditory and somatosensory) states for each lexical item. This organization of the model is straight forward with respect to findings that lexical items may be directly encoded with respect to sensory and motor representations (Coleman 1999, Roy et al. 2008, Aziz-Sadeh and Damasio 2008).

While both self-organizing maps (S-Map and P-Map) and the synaptic link weights towards the domain-specific state maps (semantic and phonemic state map as well as motor plan, and sensory maps) are part of the *long-term memory* (knowledge repository), the domain-specific state maps themselves are part of the *short-term memory* (see below). By activating a specific single state (or neuron), representing a lexical item (word) within the S-Map and the word's syllables within the P-MAP, specific and typically complex state (or neural) activation patterns arise within each state map, representing the current phonemic and/or semantic state of that word and its syllables.

Moreover both self-organizing maps (S-Map and P-Map) are associatively interconnected in a bidirectional way in order to enable an association between semantic, phonological, motor plan and sensory map activations for each lexical item. Thus production starts with an activation pattern within the semantic map describing the semantic state of a lexical item, leading to a local (or single-neuron) co-activation within the S-Map. Consequently a local co-activation occurs within the P-Map, leading to a further complex co-activation pattern for the phonemic, motor plan, auditory and somatosensory states representing the syllables of that lexical item. In contrast, perception and comprehension starts from an auditory state representation which directly leads to an activation of a phonemic state (ventral pathway, see Hickok and Poeppel 2007). This furthermore leads to a local S-Map co-activation, and then results in the co-activation of a semantic state within the semantic state map, representing the meaning of a word. If perception takes place under difficult conditions (e.g. noisy environment) the dorsal pathway may be co-activated as well (ibid.; see also next paragraph). In this case the auditory state co-activates P-Map states and these P-Map states co-activate an S-Map state via the bidirectional connection of both self-organizing maps (Fig. 2).

Processing of sub-symbolic states (i.e. auditory, somatosensory, motor plan states) arises around the speech specific sensorimotor knowledge repository, called *sensorimotor knowledge repository* or *action repository*; called mental syllabary in terms of Levelt et al. (1999). Following Kröger et al. (2009) it can be assumed that the action repository comprises a central self-organizing map which is called phonetic map (P-Map). This self-organizing map is assumed to be located in a *hyper-* or *supramodal* brain region (i.e. beyond *unimodal* brain regions). But this self-organizing map is interconnected in a bidirectional way with three sub-symbolic unimodal (i.e. domain specific) state maps, i.e. the auditory state map, the somatosensory state map, and the motor plan state map, as well as with one symbolic state map, i.e. the phonemic state map. In parallel to the organization of the mental lexicon, this central self-organizing map and its links towards all domain-specific state maps are part of the long-term memory (knowledge repository), while the domain-specific state maps themselves are part of the short-term memory. A local P-Map activation leads to specific neural activation patterns for auditory, somatosensory, and/or motor plan states, which arise within the domain-specific state maps. It can be assumed that the phonemic state representation is related to the motor plan. Each syllable or word is represented here by a symbolic description of all vocal tract actions realizing that speech item, i.e. by a list of distinctive features representing each action. The organization of each syllable in onset and rhyme and the organization of these syllable constituents in segments are implicitly given by the temporal organization of the speech or vocal tract actions constituting a syllable (Kröger and Birkholz 2009).

Articulation starts with a local activation within the P-Map which results from the activation of a lexical item (via the S-Map). This leads to a co-activation of specific neural activation patterns, representing the auditory state, the somatosensory, and the motor plan state for that syllable or word. The activation of the auditory and somatosensory state means that the model now "knows" how the auditory result of the articulation process should sound, and how the articulation of the syllable or word should feel. Thus these sensory states are also called *inner or internal sensory states* and these states are important for monitoring the syllable articulation as well as the whole word production process. A typical design for a neural state map representing vocal tract actions scores (i.e. speech motor plans) is exemplified in Kröger, Birkholz et al. (2010). A speech motor plan typically represents and specifies the types of elementary movement actions (e.g. labial, apical, dorsal, full-closing, near-closing etc.), the duration and velocity (or rapidity) of each action (Kröger and Birkholz 2007), as well as the timing between all actions needed in order to build up a syllable or word. Articulation proceeds from the motor plan state towards a subsequent neuromuscular programming and execution of a succession of temporarily overlapping vocal tract actions as defined by the motor plan (also called gestural score or vocal tract action score, Kröger and Birkholz 2007).

Perception starts with peripheral to central processing of sensory signals by using peripheral sensory organs, i.e. ears, tactile sensors of the skin, and proprioceptive muscular and joint sensors. It has been shown that the articulation-perception loop (Fig. 2) is an important vehicle for learning or training sensorimotor patterns (i.e. actions) by perceiving and imitating actions produced by others and by monitoring the reproduction of these patterns by the model itself (Kröger et al. 2009). The articulation of an action or of a score of actions representing a whole syllable or word will be accepted if the comparison between the internal auditory state already learned from an external speaker and the external auditory state produced by the articulation of the model itself (i.e. resulting from self-perception) is sufficiently small. After that learning or training period, auditory perception of speech results in a co-activation of specific neurons of the P-Map. That directly leads to a co-activation of the phonemic representation of the lexical item and to a co-activation of its semantic representation via S-Map. Since this way may in addition lead to a co-activation of motor plan states via the P-Map, this perceptual path is also called the *dorsal stream* or *dorsal pathway* (Hickok and Poeppel 2007). A second more "passive" perceptual pathway is described in literature, i.e. the *ventral stream* or *ventral pathway* (ibid.), which connects neural auditory representations of an external speech signal with phonemic representations via the phonological processing module (see above).

Last but not least it should be stated that – despite the fact that the semantic state map represents high level conceptual information – this information may be located in domain-specific brain areas representing specific perceptual and/or specific motor imageries concerning that (non-abstract) object or action. Thus the semantic state map can be assumed to be widely distributed over different brain regions (Patterson et al. 2007). Moreover it can be assumed that the activation of concepts represented within the self-organizing S-Map leads to a co-activation of higher-level as well as lower-level inner or internal sensory and motor representations which are closely related with these symbolic concepts. This organization of activation is comparable to the activation of internal auditory and motor state representations of syllables as initiated by a P-Map activation for speech production, but the activation of sensory and motor states resulting from a S-Map activation co-occurs with many different kinds of cognitive activities like thinking.

Concerning the processing modules for semantic and phonological processing it is important to state that these two processing modules are not just interconnected with the S-Map or P-Map but are also directly connected with sensory processing modules in the case of the phonological map (e.g. with auditory processing in the case of the ventral route of speech perception as indicated in Fig. 2 and with visual processing for reading, not indicated in Fig. 2) and directly connected with sensory and motor processing modules in the case of semantic processing as described above.

It is very important to separate different state (or neural) activation patterns appearing in the two *self-organizing maps* introduced above (i.e. within the long-term memory) from those which appear in the *domain-specific state maps* (i.e. within the short-term memory). A specific state within the long-term memory (i.e. an item which is activated within a self-organizing map, e.g. a specific lexical item activated within the S-Map; a specific syllable, activated within the P-Map) is represented within these self-organizing maps by a *local activation pattern* (i.e. by a single neuron or locally connected neuron cluster). Thus local activation patterns represent specific *symbolic states* within the S-Map or *supramodal sub-symbolic* states within the P-Map with-in our long term memory. In contrast in the case of state representations within unimodal domain-specific state maps (e.g. semantic state, phonemic state, auditory state, somatosensory state, motor plan state map), on the one hand, each state map comprises an ensemble of spatially closely connected model neurons (as is also the case for all self-organizing maps), but on the other hand the activation pattern for a unimodal domain-specific state is *spatially distributed over the whole cortical region defined by that domain-specific state map*. Thus the representation or activation pattern of a motor plan state within the motor plan state map can be assumed to be a direct representation of an action score (Fig. 1). The neural representation or neural activation pattern of an auditory state within the auditory state map can be assumed to be a direct representation of an acoustic spectrogram, where one dimension represents bark scaled frequency and the other dimension represents time. In a comparable way the neural representation or neural activation pattern of a somatosensory state within the somatosensory state map should comprise a two-dimensional "cast" of the tactile pattern – where one dimension represents different oral regions (labial, palatal, velar, apical, pre- and postdorsal) and where the second dimension represents the time – and a "cast" of the proprioceptive pattern of different muscles and joints of lips, tongue tip, tongue body, and lower jaw. The *knowledge* of how to activate these domain-specific neural states is stored in the long term memory, i.e. within the *links* connecting specific loci of a self-organizing map (S-Map or P-Map) with a whole domain-specific state map, while the domain-specific activation patterns only arise for a short time window within each domain-specific neural state map. Thus, the domain-specific patterns can be activated *internally* from specific loci of the self-organizing maps or *externally* from a domain-specific (external) sensory excitation (Fig. 2).

6. Training the brain: knowledge acquisition

While a blueprint for the *structure* of a control module has been outlined above, it is the goal of this chapter to describe how speech *knowledge* could be acquired, i.e. how the knowledge repositories emerge during speech acquisition. It can be assumed that mainly unsupervised associative learning takes place here. While sub-symbolic state maps are "pre-wired" to peripheral processing modules and thus while sub-symbolic state representations directly result from their domain-specific peripheral processing (e.g. action score as motor plan representation, spectrogram as auditory short term representation, see Kröger, Birkholz et al. 2010), higher-level neural representations, as occurring in the supramodal P-Map and in the cognitive S-Map emerge during learning by principles of self-organization (cf. Dehaene-Lambertz et al. 2008). Simple self-organizing Kohonen networks (Kohonen 2001) can be used (Kröger et al. 2009), while more complex approaches may include more neurobiological reality (e.g. recurrent neural network approaches, e.g. Li et al. 2008). Specific sub-modules within the higher-level part of the control module (in human analogy: specific cortical brain regions), i.e. the P-Map and the S-Map are assumed to acquire the sensorimotor and semantic knowledge, but the detailed emergence and growth processes of these maps result from (individual) learning.

A basic question for starting modeling speech acquisition is: What is the driving force for a newborn to learn to speak? One reason may be that survival is better guaranteed if knowledge for allowing the subject to participate in communications is acquired; group activities guarantee survival (Fehr et al. 2002). It is important for each human subject to become capable to comprehend the intention of others, i.e. the information another person wants to communicate and to become capable to communicate his/her own intentions or messages. Thus it can be assumed that the will to communicate is innate and this will or driving force should be manifest in the brain model of communicative robots. Thus the robot always should be willing to react on a perceived action of the communication partner by using communicative actions. A further question is: What is the driving force for being willing to incur the efforts of learning to produce and to comprehend *speech*? A hypothetical answer is that the newborn in its first communication scenarios with its caretaker immediately notices that communicative manual gestures (as well as communicative facial expressions) which are produced by caretaker (i.e. by the communication partner) are accompanied by acoustic signals (i.e. by a speech signal). The newborn immediately becomes aware that the speech signal is a part of the communicative intention of the caretaker (Tomasello 2000). Thus, early speech acquisition is closely related to face-to-face communication; e.g. it has been shown that it is not possible to learn to speak just by passively watching TV; thus speech acquisition needs communication and communicative interaction (Kuhl 2004). And since speech is produced by movements of speech organs (vocal tract actions), speech can be acquired by imitation of vocal tract actions of a caretaker occurring during face-to-face communication in a comparable way as co-verbal manual actions and co-verbal facial actions are acquired (Özçalışkan and Goldin-Meadow 2005, Rizzolatti 2005).

In the case of speech a relatively complex question is: How is the toddler capable of segmenting the continuous stream of the acoustic speech signal, e.g. and utterance as basic speech unit into meaningful parts (e.g. words)? The only input a child receives is the continuous auditory signal stream of an utterance beside contextual information (i.e. concerning the contextual situation of the current communication) and beside a signal stream of eventually co-occurring manual gestures (e.g. if the caretaker points on an object) and eventually co-occurring facial expressions of the communication partner (e.g. a smiling face). This contextual information as well as the information concerning co-occurring manual and facial gestures is important: For example the production of single word utterances (or sentences always starting with "that is a . . . ") together with a manual pointing gesture towards a visible object (e.g. chair, table, window) or together with a manual gesture of presenting an object by holding it in the hand (e.g. puppet, bottle, cloth) may be a very helpful communication process for learning non-abstract nouns; similar learning or acquisition scenarios are described by Brandl (2009) and Vaz et al. (2009).

In our hypothetical model for speech acquisition two basic learning phases can be separated, i.e. the *babbling* and the *imitation* phase. During babbling the toddler produces random vocal tract actions leading to phonation-like states, proto-vocalic, and proto-syllabic states, e.g. like [bababa], see Kröger et al. (2009); i.e. during babbling the toddler produces a series of motor and sensory states which are associated with each other. Thus, during babbling the sensorimotor part of the P-Map, i.e. the links between P-Map, motor and sensory state maps emerge (Fig. 2). If sensorimotor learning has built-up the P-Map to a certain degree during babbling, the toddler is capable of starting to imitate external acoustic signals, e.g. words which are produced by communication partners (e.g. the caretaker). This is possible now, since the toddler already has trained elementary sensory-to-motor relations. This imitation training leads to a further development of the sensorimotor part of the P-Map but now in addition associations emerge between P-Map and the S-Map representing the semantic states of the word.

Thus imitation training of a communicative robot should start with training of non-abstract nouns, which are presented to the robot via a *triadic* face-to-face communication event, i.e. the *caretaker* points to or holds an *object* in his hand and says "puppet", while the *robot or toddler* understands the communicative intention of the caretaker and looks at the object and tries to imitate the words and says e.g. [pu:pu:]. This naming may be rewarded by the caretaker by a smile accompanied by a second utterance: "Yes, a puppet". Thus during this imitation training the robot or toddler learns to associate the acoustic realization, the motor realization, and the semantic feature description of a word. This kind of speech acquisition training should be done for all words needed in the communication scenarios, the robot is designed for.

Babbling and imitation training results in the emergence of the (self-organized) S-Map, representing the trained lexical items on a semantic level, capable of co-activating the semantic states (i.e. the set of semantic features) representing these words or lexical items, as well as in the emergence of a language-specific P-Map, representing all syllables of these lexical items. A neuron activation within the P-Map leads to co-activation of motor plan states, of somatosensory states, and of auditory states for each syllable. Furthermore it can be shown that babbling allows the association of sensory and motor information of proto-syllables and that babbling leads to an ordering of these proto-syllables with respect to *supramodal phonetic features*. This is exemplified for vocalic features like "front-back" and "high-low" (Kröger et al. 2009) and for consonantal features like "place of articulation" in the case of voices plosives (ibid.). But in the same way during babbling training any other phonetic feature (i.e. any other phonetic dimension) can be learned (e.g. voicing vs. voiceless, place and manner for fricatives, etc.). Thus a phonetic ordering is established in already in the prelinguistic versions of the P-Map, which are trained during babbling training (ibid.). It is also exemplified in our preliminary modeling experiments (ibid.) that categorization takes place on the supramodal phonetic space within the P-Map if subsequently language specific training (imitation training) takes place (ibid.). *Phonemic* categorization processes over *phonetic* dimensions are also postulated in exemplar theory (Pierrehumbert 2003).

For a complete babbling training, different sets of training items should be defined reflecting the naturally occurring babbling processes. These babbling training sets should be capable of elucidating the relationship between (i) motor plan and somatosensory states, reflecting the articulation and (ii) auditory states, reflecting the acoustic signal which results from articulating a specific motor plan. Different training sets need to be built for emerging the phonetic dimensions or contrasts within the P-Map: (i) a proto-vocalic training set for emerging the phonetic dimensions front-back, high-low, and rounded-unrounded (Kröger et al. 2009), (ii) a proto-place training set for emerging the phonetic dimension place of articulation (e.g. labial, apical, dorsal, Kröger et al. 2009), (iii) a proto-constriction training set for emerging the phonetic dimension manner of articulation (e.g. full closure, critical closure, central closure with lateral opening, approximant closure), (iv) a proto-voicing training set for emerging the phonetic dimension voiced-voiceless, (v) a proto-velopharyngeal training set for emerging the phonetic dimension nasal-oral. The resulting self-organizing pre-linguistic P-Map is the basis for imitation and thus for learning lexical items. Now the question concerning a further segmentation of the acoustic signal beyond words (i.e. with respect to speech sounds) and concerning the emergence of phonemic categories during imitation training can be answered. During babbling as well as during imitation training, specific portions of the acoustic signal can be associated with specific vocal tract actions; e.g. an acoustic signal gap and the preceding and following formant transitions can be associated with a labial and/or dorsal closing action (e.g. in "pin" vs. "kin" as well as in "pin" vs. "nip"). This allows the categorization of segments, e.g. as labial or dorsal, as well as to identify segment boundaries, e.g. the acoustic realization of a syllable-initial and syllable-final /p/ as in "pin" vs. "nip". Together with the awareness that different words represent different concepts (i.e. the association towards the S-Map), this allows an assembly of the phonological system of the target language under acquisition.

7. Discussion

A blueprint for a biologically plausible "brain model" for communicative robots or communicative agents as well as for the organization of basic behavioral scenarios for acquisition of speech knowledge were outlined in this paper on the basis of current literature. It has been illustrated that natural speech acquisition mainly results from learning during face-to-face communication situations. Moreover it has been argued that learning to speak is based on human-robot face-to-face communication situations, where the human acts like a caretaker or teacher and where the robot acts like a speech-acquiring toddler. This is assumed to be a fruitful basic scenario not only for learning to speak, but also for learning to communicate including the acquisition of co-verbal manual gestures, the acquisition of co-verbal facial expressions, as well as to learn to guide or to participate in more complex face-to-face communication processes. A blueprint for a brain model introduced here has been outlined in particular for speech (i.e. vocal tract actions), but can be generalized in a straightforward way for processing manual and facial communicative actions. The control module comprising the mental lexicon can be interpreted as a word lexicon, but also as a gesture lexicon (e.g. Kipp et al. 2007) or as a lexicon for facial expressions (Pelachaud and Poggi 2002), while the sensorimotor action repository can be interpreted as a vocal tract, manual, or facial action repository; see also the unified approach for communicative actions described by Kröger and Kopp et al. (2010).

It is beyond the scope of this paper to describe the acquisition of general communication behavior like how to guide or how to act and react within a complex face-to-face communication process, i.e. how to initiate complex utterances accompanied by manual gesturing and facial expressions and how to react on actions if produced by the interlocutor. But it has been illustrated that basic face-to-face communication scenarios – as they occur between a toddler and the caretaker – are initial scenarios for learning this general communication behavior. Thus a main hypothesis of this paper is that "natural" robot-human face-to-face communication only can emerge if a robot undergoes basic face-to-face communication processes as they occur with toddlers and their caretakers.

Visual recognition and identification of objects (e.g. a puppet) is an essential process during speech acquisition in order to label objects semantically (e.g. to assign semantic features like: has a face, arms, legs, can walk, feels cuddly, looks like a human but smaller, etc.); but these topics are beyond the scope of this paper and have already been addressed and partly solved in other research groups (e.g. Li et al. 2004, Plebe et al. 2010). Furthermore it is unclear whether neural network approaches are the most suited approaches for controlling communicative robots, but in seems at least reasonable to organize the control module of these robots in a brain-like manner in order to be capable of using associative unsupervised learning which directly leads to an organization of that knowledge in a self-organizing and adaptive way.

At least three processing modes of the robot can be postulated: training, production, and perception. And these three modes are interconnected with each other: On the one hand the description of training as given above indicates that training starts with perception and needs production as a part of the babbling and imitation process. On the other hand, each perception and production process over lifetime leads to new "input" and thus can be used for further learning. Furthermore the detailed description of the imitation training scenario given above indicates that imitation may be rewarded in the case of a proper imitation of a word. Thus imitation training can be seen as reinforcement training or as a training in which the training may be partly guided by the caretaker. A second type of "guidance" occurs in babbling training. Since it is not efficient to babble all possible motor plan constellations, which at least causes an unlimited training set, and since babbling phase and imitation phase overlap in time during speech acquisition, babbling can profit from imitation in a way that babbling prefers motor items which are similar to target language specific motor patterns. Thus babbling more and more becomes language specific within the first year of lifetime (Goldstein and Schwade 2008, Kuhl 2004).

It is an important feature of the hypothetical brain model introduced here to separate lower-level and higher-level processing. Higher-level cognitive processes are stimulated by internal or inner representations (imagery) of percepts or actions (i.e. lower-level inner representations) and mainly process symbolic representations, which are associated with these sensory or motor imageries and which represent the meaning of these lower-level representations (Haikonen 2009, p.46ff). These symbolic representations are effective processing units since symbolic representations are more "compressed"; i.e. only a brief representation is needed to be activated in the case of symbolic states in comparison to perceptual or motor representations. Thus higher-level symbolic representations can be labeled as "compressed" or brief representations and these representations disburden the brain and allow a widening of the time window for conceptualization and planning of complete sentences or utterances, since the capacity of the short-term working memory is limited. While a temporal processing interval on the sensorimotor level comprises only few syllables, the temporal processing interval on the semantic level comprises complete sentences or utterances (for a discussion of different time scales in cortical and subcortical processing see Kiebel et al. 2008).

Last but not least it will be shown that the blueprint of a brain model introduced in this paper (Fig. 2) is well motivated from a neurobiological viewpoint, since all modules and maps defined in this hypothetical model can be located anatomically in real brains. Starting with articulation, the motor plan map – hosting neural presentations of currently active motor plan states – is assumed to be located in the premotor cortex and/or in the supplementary motor area SMA (Riecker et al. 2005). Neuromuscular programming is assumed to be hosted here as well as in subcortical structures (e.g. cerebellum, parts of the basal ganglia, ibid.). Execution starts on the level of the primary motor cortex and proceeds via subcortical structures towards the peripheral neuromuscular units directly controlling the movements of the vocal tract articulators. Somatosensory processing starts at tactile and proprioceptive receptor cells within the vocal tract and proceeds via subcortical structures (e.g. thalamus) towards primary and higher unimodal somatosensory cortical regions which are located in the anterior inferior parietal lobe (Kandel et al. 2000). Auditory processing starts at auditory receptor cells within the inner ear and proceeds via subcortical structures (e.g. thalamus) towards primary and higher unimodal unilateral auditory cortical regions which are located in the dorsal superior temporal gyrus (ibid.). While the motor plan state map is located in the premotor and/or supplementary motor area of the frontal lobe, the somatosensory state maps for processing internal as well as external somatosensory states are located in the anterior inferior parietal lobe (i.e. a part of the parietal lobe) and the auditory state maps for processing internal as well as external auditory states are located in the dorsal superior temporal gyrus (i.e. a part of the temporal lobe). Thus it can be seen that these unimodal domain-specific state maps which are related to the motor and different sensory domains are well separated in the brain in three of four different cortical lobes; moreover visual state maps are located in the fourth, i.e. in the occipital lobe.

The anatomical location of the neural maps and processing modules representing higher-level symbolic or cognitive states is less specific. It can be stated that the phonological processing module as well as the phonemic state map is located bilaterally in the mid-post superior temporal gyrus (mid-post STS, Hickok and Poeppel 2007) while the hyper- or supramodal P-Map is assumed to be located in the posterior middle and inferior portions of both temporal lobes with a weak left-hemisphere bias (i.e. lexical interface, ibid.). The semantic state map as well as the semantic processing module represent a neural network which is widely distributed over the whole cerebral cortex, including the anterior temporal cortex (basic combinatorics and semantic integration with context, Lau et al. 2008) and including anterior and posterior portions of the inferior frontal cortex for controlled retrieval and selection of lexical items (ibid.). The S-Map which connects all domain-specific sensory and motor semantic state representations (semantic map) can be compared to a supramodal semantic hub, which is assumed to be located in the anterior temporal lobes (Patterson et al. 2007).

It is the main goal of this paper to inspire robot constructing engineers to develop control modules as well as to design the learning or training scenarios for future exemplars of humanoid face-to-face communication robots in the way that is described in this paper. Modeling not only the visual shape of a robot in a human-like way, but also its control structures as well as its knowledge acquisition as natural as possible, may in principle overcome theoretical and practical limits occurring for naturalness of robot acting and reacting, i.e. limits in action perception and action recognition as well as limits in action initiation and action production as they occur in currently available artificial systems which are not designed with respect to principles of neurobiology.

Acknowledgments

This work was supported in part by German Research Council (DFG) grant Kr 1439/13-1 and grant Kr 1439/15-1 and in part by COST-action 2102.

References

Asada M, Hosoda K, Kuniyoshi Y, Ishiguro H, Inui T, Yoshikawa Y, Ogino M, Yoshida C, 2009. Cognitive developmental robotics: A survey. IEEE transactions on Autonomous Mental Development 1, 12-34.

Aziz-Sadeh L, Damasio A, 2008. Embodied semantics for actions: Findings from functional brain imaging. Journal of Physiology-Paris 102, 35-39.

Bailly G, Raidt S, Elisei F, 2010. Gaze, conversational agents and face-to-face communication. Speech Communication 52, 598-612.

Bergmann K, Kopp S, 2009. Increasing the Expressiveness of Virtual Agents – Autonomous Generation of Speech and Gesture for Spatial Description Tasks. In: Decker K, Sichman J, Sierra C, Castelfranchi C (eds.) Proceedings of the 8^{th} International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), pp. 361-368.

Birkholz P, Kröger BJ, 2006. Vocal tract model adaptation using magnetic resonance imaging. Proceedings of the 7^{th} International Seminar on Speech Production (Belo Horizonte, Brazil) pp. 493-500.

Birkholz P, Kröger BJ, Neuschaefer-Rube C, in press. Model-based reproduction of articulatory trajectories for consonant-vowel sequences. IEEE Transactions on Audio, Speech and Language Processing. DOI:10.1109/TASL.2010.2091632

Brandl H, 2009. A computational model for unsupervised child-like speech acquisition. Unpublished Doctoral Thesis (University of Bielefeld, Bielefeld, Germany)

Breazeal C, 2003. Towards sociable robots. Robotics and Autonomous Systems 42, 167-175.

Breazeal C, 2004. Function meets style: Insights from emotion theory applied to HRI. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 34, 187-194.

Brooks RA, Breazeal C, Marjanovic M, Scassellati B, Williamson MM, 1999. The cog project: building a humanoid robot. In: Nehaniv CL (ed.) Computation for metaphors, analogy, and agents (Springer Verlag, Berlin), pp. 52-87.

Caligiore D, Ferrauto T, Parisi D, Accornero N, Capozza M, Baldassarre G, 2008. Using motor babbling and Hebb rules for modeling the development of reaching with obstacles and grasping. In: Dillmann R, Maloney C, Sandini G, Asfour T, Cheng G, Metta G, Ude A (eds.) International Conference on Cognitive Systems, CogSys2008 (University of Karlsruhe, Karlsruhe, Germany)

Cangelosi A, Riga T, 2006. An embodied model for sensorimotor grounding and grounding transfer: experiments with epigenetic robots. Cognitive Science 30, 673-689.

Coleman J, 1999. Cognitive reality and the phonological lexicon: A review. Journal of Neurolinguistics 11, 295-320.

Dehaene-Lambertz G, Hertz-Pannier L, Dubois J, Dehaene S, 2008. How Does Early Brain Organization Promote Language Acquisition in Humans? European Review 16, 399-411.

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), pp. 31-37.

Desmurget M, Grafton ST, 2000. Forward modeling allows feedback control for fast reaching movements. Trends in Cognitive Sciences 4, 423-431. Dohen M, Schwartz, JL, Bailly G, 2010. Speech and face-to-face communication – An introduction. Speech Communication 52, 477-480.

Fehr E, Fischbacher U, Gächter S, 2002. Strong reciprocity, human cooperation, and the enforcement of social norms. Human Nature 13, 1-25.

Fujie S, Fukushima K, Kobayashi T, 2004. A conversation robot with backchanel feedback function based on linguistic and non-linguistic information. Proceedings of the 2^{nd} International conference on Autonomous Robots and Agents (Palmerston North, New Zealand), pp. 379-384.

Fukui K, Nishikawa K, Ikeo S, Shintaku E, Takada K, Takanobu H, Honda M, Takanishi A, 2005. Development of a talking robot with vocal cords and lips having human-like biological structures. Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (Edmonton, Alberta, Canada), pp. 2023-2028.

Galantucci B, Steels L, 2008. The embodied communication in artificial agents and humans. In: Wachsmuth I, Lenzen M, Knoblich G (eds.), Embodied Communication in Humans and Machines (Oxford University Press, Oxford) pp. 229-256.

Goldstein MH, Schwade J, 2008. Social Feedback to Infants' Babbling Facilitates Rapid Phonological Learning. Psychological Science 19, 515-523.

Goldstein MH, Schwade J, Briesch J, Syal S, 2010. Learning While Babbling: Prelinguistic Object-Directed Vocalizations Indicate a Readiness to Learn. Infancy 15, 362-391.

Golfinopoulos E, Tourville JA, Bohland JW, Ghosh SS, Nieto-Castanon A, Guenther FH, 2011. fMRI investigation of unexpected somatosensory feedback perturbation during speech. NeuroImage 55, 1324-1338.

Grossberg S, 2010. The link between brain learning, attention, and consciousness. In: Carsetti A (ed.) Causality, Meaningful Complexity and Embodied Cognition (Springer, Dordrecht), pp. 3-45.

Grossmann T, Johnson MH, Lloyd-Fox S, Blasi A, Deligianni F, Elwell C, Csibra G, 2008. Early cortical specialization for face-to-face communication in human infants. Proceedings of the Royal Society B: Biological Sciences 275, 2803-2811.

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

Haikonen POA, 2009. The role of associative processing in cognitive computing. Cognitive Computation 1, 42-49.

Hashimoto T, Kato N, Kobayashi H, 2010. Study on educational application of android robot SAYA: Field trial and evaluation at elementary school. In: Lui H, Ding H, Xiong Z, Zhu X (eds.) Intelligent Robotics and Applications. LNCS 6425 (Springer, Berlin), pp. 505-516.

Hickok G, Poeppel D, 2007. Towards a functional neuroanatomy of speech perception. Trends in Cognitive Sciences 4, 131–138.

Indefrey P, Levelt WJM, 2004. The spatial and temporal signatures of word production components. Cognition 92, 101-144.

Iverson JM, Capirci O, Longobardi E, Caselli MC, 1999. Gesturing in mother-child interactions. Cognitive Development 14, 57-75.

Kanda T, Hirano T, Eaton D, 2004. Interactive robots as social partners and peer tutors for children: a field trial. Human-Computer Interaction 19, 61-84.

Kanda T, Kamasima M, Imai M, Ono T, Sakamoto D, Ishiguro H, Anzai Y, 2007. A humanoid robot that pretends to listen to route guidance from a human. Journal of Autonomous Robots 22, 87-100.

Kanda T, Miyashita T, Osada T, Haikawa Y, Ishiguro H, 2008. Analysis of humanoid appearance in human-robot interaction. IEEE Transactions on Robotics 24, 725-735.

Kandel ER, Schwartz JH, Jessell TM, 2000. Principles of Neural Science. 4^{th} edition (McGraw-Hill, New York).

Kiebel SJ, Daunizeau J, Friston KJ, 2008. A Hierarchy of Time-Scales and the Brain. PLoS Comput Biol 4(11): e1000209. doi:10.1371/journal.pcbi.1000209.

Kipp M, Neff M, Kipp KH, Albrecht I, 2007. Towards Natural Gesture Synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis. In: Pellachaud C, Martin JC, Andre E, Chollet G, Karpouzis K, Pele D (eds.), Intelligent Virtual Agents. LNAI 4722 (Springer, Berlin), pp. 15-28.

Kohonen T, 2001. Self-Organizing Maps (Springer, Berlin).
Kopp S, Bergmann K, Buschmeier H, Sadeghipour A, 2009. Requirements and Building Blocks for Sociable Embodied Agents. In: Mertsching B, Hund M, Aziz Z (eds.) Advances in Artificial Intelligence. LNCS 5803 (Springer, Berlin), pp. 508-515.
Kopp S, Gesellensetter L, Krämer NC, Wachsmuth I, 2005. A Conversational Agent as Museum Guide – Design and Evaluation of a Real-World Application. In: Panayiotopoulos T, Gratch J, Aylett R, Ballin D, Oliver P, Rist T (eds.), Intelligent Virtual Agents. LNCS 3661 (Springer, Berlin), pp. 329-343.
Kosuge K, Hirata Y, 2004. Human-robot interaction. Proceedings of the 2004 IEEE International Conference on Robotics and Biometrics (Xhenyang, China), pp. 8-11.
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, Birkholz P, 2009. Articulatory Synthesis of Speech and Singing: State of the Art and Suggestions for Future Research. In: Esposito A, Hussain A, Marinaro M (eds) Multimodal Signals: Cognitive and Algorithmic Issues. LNAI 5398 (Springer, Berlin), pp. 306-319.
Kröger BJ, Kannampuzha J, Neuschaefer-Rube C, 2009. Towards a neurocomputational model of speech production and perception. Speech Communication 51, 793-809.
Kröger BJ, Birkholz P, Lowit A, 2010. Phonemic, sensory, and motor representations in an action-based neurocomputational model of speech production (ACT). In: Maassen B, van Lieshout P (eds.), Speech Motor Control: New developments in basic and applied research. (Oxford University Press, New York), pp. 23-36.
Kröger BJ, Kopp S, Lowit A, 2010. A model for production, perception, and acquisition of actions in face-to-face communication. Cognitive Processing 11, 187-205.
Kuhl PK, 2004. Early language acquisition: cracking the speech code. Nature Reviews Neuroscience 5, 831-843.
Kuhl PK, 2007. Is speech learning „gated by the social brain? Developmental Science 10, 110-120.
Lau EF, Phillips C, Poeppel D, 2008. A cortical network for semantics: (de)constructing the N400. Nature Reviews Neuroscience 9, 920-933.
Levelt WJM, Roelofs A, Meyer A, 1999. A theory of lexical access in speech production. Behavioral and Brain Sciences 22, 1-75.
Li P, Fakas I, MacWhinney B, 2004. Early lexical development in a self-organizing neural network. Neural Networks 17, 1345-1362.
Li Y, Kurata S, Morita S, Shimizu S, Munetaka D, Nara S, 2008. Application of chaotic dynamics in a recurrent neural network to control: hardware implementation into a novel autonomous roving robot. Biological Cybernetics 99, 185-196.
Lindblom J, Ziemke T, 2003. Social situatedness of natural and artificial intelligence: Vygotsky and beyond. Adaptive Behavior 11, 79-96.
Lungarella M, Metta G, Pfeiffer R, Sandini, 2003. Developmental robotics: a survey. Connection Science 15, 151-190.
Madden C, Hoen M, Dominey PF, 2010. A cognitive neuroscience perspective on embodied language for human-robot cooperation. Brain and Language 112, 180-188.
McGurk H, MacDonald J, 1976. Hearing lips and seeing voices. Nature 264, 746-748.
Mitchell CJ, De Houwer J, Lovibond PF, 2009. The propositional nature of human associative learning. Behavioral and Brain Sciences 32, 183-198. Ogawa H, Watanabe T, 2000. Interrobot: A speech driven embodied interaction robot. Proceedings of the 2000 IEEE International Workshop on Robot and Human Interactive Communication (Osaka, Japan), pp. 322-327.
Özçalkan S, Goldin-Meadow S, 2005. Gesture is at the cutting edge of early language development. Cognition 96, B101-B113.
Parisi D, 2010. Robots with language. Frontiers in Neurorobotics 4. DOI: 10.3389/fnbot.2010.00010
Patterson K, Nestor PJ, Rogers TT, 2007. Where do you know what you know? The representation of semantic knowledge in the human brain. Nature Reviews Neuroscience 8, 976-987.
Pelachaud C, Poggi I, 2002. Subtleties of facial expressions in embodied agents. The Journal of Visualization and Computer Animation 13, 301–312. Pierrehumbert JB, 2003. Phonetic diversity, statistical learning, and acquisition of phonology. Language and Speech 46, 115-154.
Plebe A, Mazzone M, de la Cruz V, 2010. First word learning: a cortical model. Cognitive Computation 2, 217-229.
Prince CG, Demiris Y, 2003. Introduction to the special issue on epigenetic robotics. Adaptive Behavior 11, 75-77.
Rich C, Ponsler B, Holroyd A, Sidner CL, 2010. Recognizing engagement in human-robot interaction. Proceedings of the 5th ACM/IEEE International conference on Human-Robot Interaction (Osaka, Japan), pp. 375-382.
Riecker A, Mathiak K, Wildgruber D, Erb A, Hertrich I, Grodd W, Ackermann H, 2005. fMRI reveals two distinct cerebral networks subserving speech motor control. Neurology 64, 700-706.
Rizzolatti G, 2005. The mirror neuron system and its function in humans. Anatomy and Embryology 210, 419-421.
Roy AC, Craighero L, Fabbri-Destro, M, Fadiga L, 2008. Phonological and lexical motor facilitation during speech listening: A transcranial magnetic stimulation study. Journal of Physiology-Paris 102, 101-105.
Saunders JA, Knill DC, 2004. Visual Feedback Control of Hand Movements. The Journal of Neuroscience 24, 3223-3234.
Schaal S, 1999. Is imitation learning the route to humanoid robots? Trends in Cognitive Sciences 3, 233-242.
Shiomi M, Kanda T, Miralles N, Miyashita T, 2004. Face-to-face interactive humanoid robot. Proceedings of the 2004 IEEE International Conference on Intelligent Robots and Systems (Sendai, Japan), pp. 1340-1346.
Shiwa T, Kanda T, Imai M, Ishiguro H, Hagita N, 2008. How quickly should communication robots respond? Proceedings of 2008 ACM Conference of Human Robot Interaction (Amsterdam, Netherlands), pp. 153-160.
Sidner CL, Lee C, Kidd CD, Lesh N, Rich C, 2005. Explorations in engagement for humans and robots. Artificial Intelligence 166, 140-164. Steels L, 2003. Evolving grounded communication for robots. Trends in Cognitive Sciences 7, 308-312.
Tani J, Ito M, 2003. Self-organization of behavioral primitives as multiple attractor dynamics: a robot experiment. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans 33, 481-488.
Tani J, Nishimoto R, Namikawa J, Ito M, 2008. Codevelopmental learning between human and humanoid robot using a dynamic neural network model. IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics 38, 43-59.
Thompson RF, 1986. The neurobiology of learning and memory. Science 233, 941-947.
Tomasello M, 2000. First steps towards a usage-based theory of language acquisition. Cognitive Linguistics 11, 61-82.
Trappenberg T, Hartono P, Rasmusson D, 2009. Top-Down Control of Learning in Biological Self-Organizing Maps. In: Principe JC, Miikkulainen R (eds.), Advances in Self-Organizing Maps. LNCS 5629 (Springer, Berlin), pp. 316-324.

Yoshikawa Y, Shinozawa K, Ishiguro H, Hagita N, Miyamoto T, 2006. The effects of responsive eye movement and blinking behavior in a communication robot. Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (Beijing, China), pp. 4564-4569.
Vaz M, Brandl H, Joublin F, Goerick C, 2009. Learning from a tutor: Embodied speech acquisition and imitation learning. Proceedings of the IEEE 8^{th} International Conference on Development and Learning (Shanghai, China), pp. 1-6.
Vilhjálmsson H, 2009. Representing communicative function and behavior in multimodal communication. In: Esposito A, Hussain A, Marinaro M, Martone R (eds.) Multimodal Signals: Cognitive and Algorithmic Issues. LNCS 5398 (Springer, Berlin), pp. 47-59.
Weng J, 2004. Developmental robotics: Theory and experiments. International Journal of Humanoid Robotics 1, 199-236.
Weng J, McClelland J, Pentland A, Sporns O, Stockman I, Sur M, Thelen E, 2001. Autonomous mental development by robots and animals. Science 291, 599-600.