Social Motorics – Towards an Embodied Basis of Social Human-Robot Interaction

Amir Sadeghipour, Ramin Yaghoubzadeh, Andreas Rüter, and Stefan Kopp

Abstract In this paper we present a biologically-inspired model for social behavior recognition and generation. Based on an unified sensorimotor representation, it integrates hierarchical motor knowledge structures, probabilistic forward models for predicting observations, and inverse models for motor learning. With a focus on hand gestures, results of initial evaluations against real-world data are presented.

1 Introduction

For human-centered robots to be able to engage in social interactions with their users, they need to master a number of daunting tasks. This includes, e.g., robust and fast recognition and understanding of interactive user behavior, human-acceptable expressivity, joint attention, or incremental dialog with mutual adaptivity. In humans such capabilities are assumed to rest upon an embodied basis of social interaction – direct interactions between perception and generation processes (*perception-behavior expressway* [6]) that support mirroring or resonance mechanisms [15] to process social behavior at different levels, from kinematic features to motor commands to intentions or goals [8]. Research in social robotics has increasingly started to adopt such principles in its work on architectures and interaction models (e.g. [3, 5]). Against this background we present our work towards "social motorics", modeling a resonant sensorimotor basis for observing and using social behavior in human-robot interaction. With a focus on hand-arm gestures, we describe a probabilistic model that exploits hierarchical motor structures with forward and inverse models in order to allow resonance-based processing of social behavior. We start in Section 2 with a review of related work on gesture learning and recognition. In Section 3 we introduce our overall computational model and detail the employed forward and inverse models in Section 4 and 5, respectively. The probabilistic modeling of interactions between perception and generation is described in Section 6 and, afterwards, we present in Section 7 results of how the model performs at simulating resonances during perception of gestural behavior. Finally, Section 8 gives a conclusion and summery of future works.

Corresp. author: A. Sadeghipour
Sociable Agents Group, CITEC, Bielefeld University e-mail: asadeghi@techfak.uni-bielefeld.de
R. Yaghoubzadeh, S. Kopp
Sociable Agents Group, CITEC, Bielefeld University

2 Related Works

The growing interest in developing social artificial agents requires abilities for perception, recognition and generation of gestures as one of the non-verbal interaction modalities. The recognition process is concerned with the analysis of spatio-temporal features of the hand movements and is mainly treated as pattern classification with subsequent attribution of meaning. Many studies apply probabilistic approaches to classify hand gesture trajectories. Hidden Markov models have been widely applied as an efficient probabilistic approach to work with sequence of data [1, 4, 7]. However, the focus of those approaches is on pattern recognition separated from the attribution of meaning, and they rely on hidden variables which do not directly correspond to the agent's own action repertoire. For recognizing transitive actions, in which the goal of an observed action is often visually inferable, hierarchical models are used to analyze the perceived stimuli in a bottom-up manner towards more abstract features and, consequently, goals of those actions [2, 11, 17].

Furthermore, imitation mechanisms (overt or covert ones) are widely used for learning and reproducing behaviors in artificial agents. For instance, the MOSAIC model applies forward and inverse models to predict and control movements in a modular manner [9]. Others [10, 17] have worked on hierarchical MOSAIC models towards more abstract levels of actions. However, none of these models has adopted imitation mechanisms to attain perception-action links using a shared motor representation – a hallway of what we assume here to be the basis of social motorics.

3 Resonant Sensorimotor Basis

In our work we aim at modeling perceptuo-motor processes that enable a robot, on the one hand, to concurrently perceive, recognize and *understand* motor acts of hand-arm gestures and learn them by imitation (cf. [12, 14]). That is, the model is to process the robot's perceptions automatically, incrementally, and hierarchically from hand and arm movement observation toward understanding and semantics of a gesture. In result, the robot's motor structures are to start to "resonante" to the observation of corresponding actions of another structurally congruent agent (either human or artificial). On the other hand, the model is designed to allow the robot to *generate* gestures in social interaction from the same motor representation.

Overall, the model connects four different structures (Fig. 1): preprocessing, motor knowledge, forward models, and inverse models. We presume that some kind of perceptual processing has identified a human's body parts as relevant for hand-arm gesture. Now, the preprocessing module receives continuous stimuli about the user's hand postures (finger configurations) and wrist positions in the user's effector space. Since in our framework the received sensory data are already associated with corresponding body parts (left/right arm and left/right hand) of the human demonstrator (cf. Section 7), and since we assume the robot to be anthropomorphic, body correspondence can be established straight-forward. A *body mapping* submodule maps

the perceived data from the human-centered coordinate system into the robot's body frame of reference. The *sensory memory* receives the transformed visual stimuli at each time step and buffers them in chronological order in a *working memory*.

The motor knowledge structures encode the robot's competence to perform certain gestural behaviors itself – more specifically, to perform the required movements of the relevant body parts. This knowledge is organized hierarchically. The lowest level contains *motor commands* (MC) required for the single movement segments (cf. [13]). Data for each of the four relevant body parts is stored in directed graphs, the nodes of which are intermediate states within a gestural movement; the edges represent the motor commands that lead from one state to another. The next level, also present for each body part, consists of *motor programs* (MP) that cluster several sequential MCs together and represent paths in the motor command graph. Each MP stands for a meaningful movement, i.e., a gestural performance executed with the respective body part. Since gestures typically employ both the hand as well as the more proximal joints (elbow, shoulder), often even in both arms, all contributing body parts need to be controlled simultaneously. Furthermore, gestures are generally not restricted to a specific performance but have some variable features which, when varied, do not change the meaning of the gesture but merely the way of performing it. Thus, a social robot must be able to cluster numerous instances of a gestural movement into a so-called "schema", which demarcates the stable, mandatory features from the variable features. For example, a waving gesture has a number of determinant features (hand lifted, palm facing away from the body, reciprocating

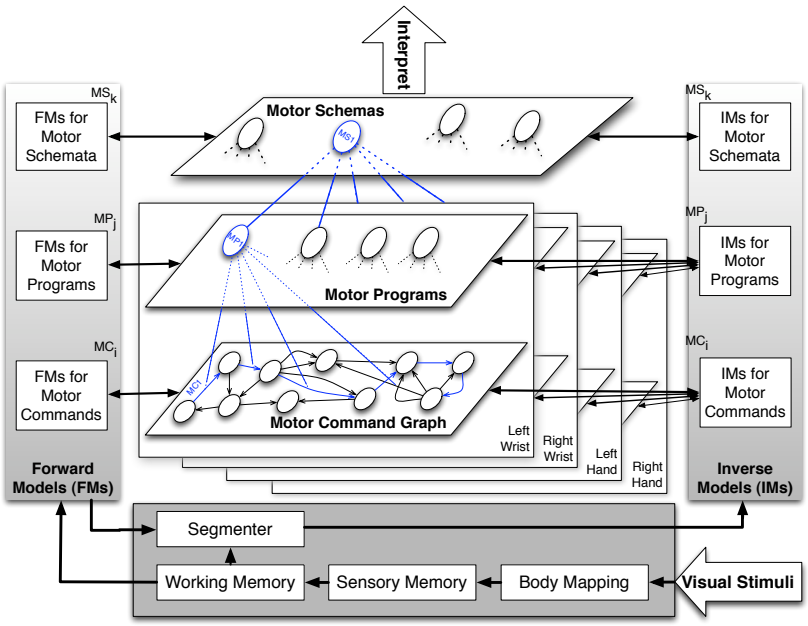

Fig. 1 Outline of the structure of a sensorimotor basis for social motorics.

motion in the frontal body plane), while bearing a number of variable features that mark its context-dependency or manner of execution (e.g., number of repetitions, speed, handedness, height). Therefore, we define *motor schemas* (MS) as a generalized representation that groups different, familiar performances in relevant body parts (MPs) of a gesture into a single cluster. Such a generalization process can foster the understanding and imitation of behavior in several ways. First, by combining different body parts into an MS, a gesture can be recognized more robustly combining information about different body parts. Second, the concept of motor schemas elevates the problem of interpreting a gesture from the complex motor level to a more abstract, yet less complex level, namely schema interpretation. Third, a robot can retain its own personal form of performing a gesture while being able to relate other performances of the same gesture to the same schema.

The third structure is formed by forward models. Such models are derived from the robot's motor knowledge at each level. While observing a behavior, they run *internal* simulations in order to predict how the behavior would continue for each possible explanation considered. By evaluating this prediction against the actual percepts at each time step, this structure is able to determine how well individual motor commands, programs or schemas correspond to the observed behavior. If there is no sufficiently corresponding representation, the processing switches to the final structure, the inverse models. These are responsible for learning, i.e. analyzing the movements of a behavior and augmenting the robot's motor knowledge at all three levels correspondingly. To this end, a *segmenter* on the lowest level decomposes received movements into (nearly) planar segments based on their kinematic features (i.e. velocity profile and direction changes).

In the following, we will present a probabilistic approach to model these three core structures (motor knowledge, forward models and inverse models) for intransitive, gestural movements.

4 Forward Models

The classification of observed input into levels of increasing abstraction, as described above, is achieved by matching it with simulations performed according to the receiver's own motor repertoire. This simulation and matching is performed by the *forward models* in a probabilistic way. The aim is to find a set of hypotheses from the repertoire that can explain the observed input. For each hypothesis considered, a class of functors termed *predictors* constructs a probability density function for the input likelihood for an arbitrary time in the future (the *prediction*), under the condition that the motor component associated with that hypothesis were to be the one producing the observations. The result of the ongoing evaluation of these expectations against the actual evidence is assumed to reflect automatic "resonances" in the robot's motor hierarchy. As demanded for an embodied basis of social interaction, this method therefore consequently carries the assumption of a correspondence between the motor repertoires of the human user and the social robot.

Based on the predictors, any observation is evaluated against the densities predicted for that time. This yields a measure of *explainability* of the data under the assumptions implied by the hypothesis (*diagnostic support*). The resulting *a-posterori* performances of the hypotheses are then compared using Bayes' theorem, taking into account the probability of certain motion primitives as provided in the form of prior distributions, which can be influenced by higher levels. The explicit posterior distributions are also used for a suitable pruning of the search space, allowing both the retention of plausible hypotheses while at the same time discarding those hypotheses deemed negligible. Full probabilistic forward models for the levels of motor commands, motor programs and motor schemas for wrist trajectories in 3D space have been implemented and tested (cf. Section 7). The forward model at the motor command level makes use of distribution functions formed by the convolution of a configurable Gaussian kernel along parts of the possible trajectories as spanned by consecutive motor commands. The covered 3D space is also a function of time, which is addressed by another configurable distribution function relating the individual tolerated speed variance to a path segment along the trajectory. This set of variable-density "tubular clouds" (Fig. 2(d)) is utilized as hypothesis-dependent likelihood functor $P(\mathbf{o}_t|c)$. Formula 1 details the generation of posteriors with a *prior feedback* approach (using the previous posterior as prior $P_{T-1}(c)$).

$$P_T(c|\mathbf{o}) := \frac{1}{T}\sum_{t=t_1}^{T} P(c|\mathbf{o}_t) = \frac{1}{T}\sum_{t=t_1}^{T} \alpha_c P_{T-1}(c) P(\mathbf{o}_t|c) \tag{1}$$

The forward models at higher levels work similarly in a Bayesian fashion and consider the observation and hypotheses from the lower levels. The motor program hypotheses (Formula 2) contain additionally the likelihood functor $P(c|p)$, which is modeled as a simple discrete Gaussian probability distribution along the according motor commands at each time step.

$$P_T(p|C,\mathbf{o}) := \frac{1}{T}\sum_{t=t_1}^{T} \alpha_p P_{T-1}(p) \sum_{c\in C} P(\mathbf{o}_t|c) P_t(c|p) \tag{2}$$

A motor schema clusters different performances of a gesture with certain variable features. The corresponding forward model at this level contains a new likelihood functor $P(p|s)$ that equals one, only if the according motor program, p, is clustered to the given motor schema, s. In addition, the forward model contains both likelihood models of the lower levels. Hence, the parameters of those likelihood functors (variances) can be set by each motor schema in order to take the variable performance features into consideration, i.e. velocity and position of motor commands or repetition of a movement segment. Furthermore, motor schemas determine which body parts contribute to the performance by applying an AND- or OR-relation (sum, product or a combination of both) to combine the prediction probabilities adequately. Formula 3 considers the case of performing a gesture using all four body parts: right/left arm (*rw*/*lw*) and right/left hand (*rh*/*lh*).

$$P_T(s|\mathbf{C},\mathbf{P},\mathbf{o}_{lw},\mathbf{o}_{rw},\mathbf{o}_{lf},\mathbf{o}_{rf}) := \frac{1}{T}\sum_{t=t_1}^{T}\alpha_s P_{T-1}(s)\prod_{i\in\{rw,lw,rh,lh\}}\sum_{p\in P}P(p_i|s)\sum_{c\in C}P(\mathbf{o}_{i,t}|c_i)P_t(c_i|p_i) \quad (3)$$

In future work, biological constraints and proprioceptive information will be applied during the simulation. This will allow a richer attribution to an observed movement (e.g. as being effortful and hence emphasized). Furthermore, the forward models will also be used to assess the feasibility of hypothetical motor commands for the robot before enriching the repertoire or executing them for (true) imitation.

5 Inverse Models

Whenever a novel behavior is observed, i.e. the forward models have failed to yield a sufficient explainability from the known repertoire, an *inverse model* takes over. It is in charge of formulating motor structures that can reconstruct the novel observation at the respective level of representation, thereby allowing for extension of the robot's repertoire. In our approach, the learning of gestures at the MC level uses a self-organizing feature map (SOM) to map observations over time onto a lower dimensional grid of neurons that represent prototypes, derived from gestures perceived in the past. These prototypes are used for classification and the generation of motor commands that form the repertoire of the robot (Fig. 2(a-c)). The best-matching neuron is determined via a "winner-takes-all" approach and its neighbourhood is adjusted by the difference between the input and the best match. The emergent map features smooth transitions between adjacent prototypes.

The inverse model for motor commands operates on movement segments. The input data are present as a sequence of nearly planar 3D segments, which are first projected into 2D using PCA, transformed into a common coordinate frame, and sampled in equidistant intervals. This normalization, which is inverted during final reconstruction, allows for the comparison of prototypes and input necessary for classification and training of the SOM. We use a randomly initialized, dynamical and online-learning SOM, which enables classification while in training. The dynamic learning process is controlled in order to prevent overfitting of the map, and eventually suspended until new input is presented.

From the winner neurons for the single movement segments, correspondingly parameterised MCs can be directly computed (cf. [13]) and imparted to the motor command graph. This will also insert a new MP, if there is none which consists of the sequence of the winner MCs. The motor schema level reaches into the context of gesture use and needs to cluster instances of gesture performance. This decision is subject of ongoing work, in which we consider how imitation with informative feedback from a human interlocutor can scaffold the learning of invariant features of a gesture schema and fitting of the likelihood parameters of the corresponding forward model at this level when new performances of familiar gestures are observed.

6 Resonance-based Behavior Processing

The proposed model employs one and the same motor knowledge to guide the recognition of familiar hand-arm gestures, and as repository of motor commands and programs in the self-generation of behavior. Such a direct link between perception and action is assumed to underlie the evident cross-activation and influence of the two processes. The resulting mirroring of actions made by another individual is assumed to be fundamental to social understanding and embodied communication [15]. Such resonances in sensorimotor structures can enable many mutualities abundant in social interaction [6], e.g. non-conscious mimicry when leaking through to execution, or alignment when leaving traces that affect behavior production.

In our model, perception-induced resonances are the posterior probabilities of valid hypotheses. It is also simulated how such resonances percolate upwards, from single motor commands to higher-level structures, as well as how higher levels may affect and guide the perception process at lower levels over the next time steps. These processes are accounted for by computing the posteriors using Bayes' law and inserting prior probabilities for each motor component which depend on three criteria: (1) the number of candidate hypotheses (assigning the default priors); (2) the a-posteriori from upper levels (cf. Section 4); (3) the posterior probability in the previous time step, since we apply the prior-feedback method to model time dependency between sequential evidences. The combination of these priors affects the activation of the corresponding motor component during perception.

Except for the first one, these criteria also carry on information about the last perceived gesture. Therefore, these priors are not directly reset to their default values after perception, but decline following a sigmoidal descent towards the default a-priori. When the robot, as advocated here, uses the same motor knowledge and consequently the same prior probabilities while selecting proper motor components for producing its own behavior, the robot tends to favor those schemas, programs, and motor commands that have been perceived last. The other way around, the model also allows to simulate "perceptual resonance": choosing a motor component for generation increases its prior probability temporarily, biasing the robot's perception toward the self-generated behavior – another suggested mechanism of coordination in social interaction [16].

7 Results

The proposed model for resonance-based gesture perception has been implemented and tested with real-world gesture data in a setup with a 3D time-of-flight camera (SwissRanger^{TM}SR4000[1]), which can be easily mounted to any mobile robot, and the marker-free tracking software iisu[2].

[1] http://www.mesa-imaging.ch

[2] http://www.softkinetic.net

Inverse model The performance of the applied SOM at the MC level depends on the training data and parameters. The result of a trained 4×4 SOM is shown in Fig. 2(a-c), after segmenting the observed wrist trajectory of a figure "3" drawn in the air.

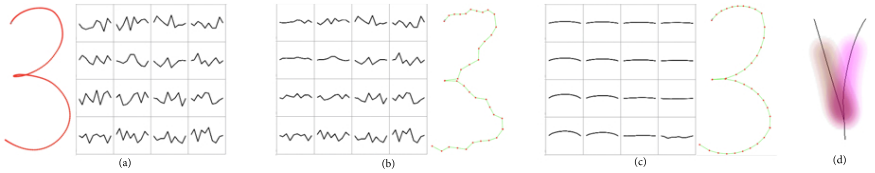

Fig. 2 (a) The performed figure "3" trajectory and a randomly initialized SOM; (b) the SOM and mapped trajectory after 10 training iterations and (c) after 150 iterations; (d) visualization of a time-dependent likelihood function $P(o_t|h)$ used by the forward models.

Forward models In the following example, the motor knowledge (Fig. 3 top-left) was built based on observation of several performances of four different gestures: waving, drawing a circle, and two variants of pointing upwards. Fig. 3 shows how the confidences of alternative hypotheses at all three motor levels are evolving *during* the perception of another waving gesture.

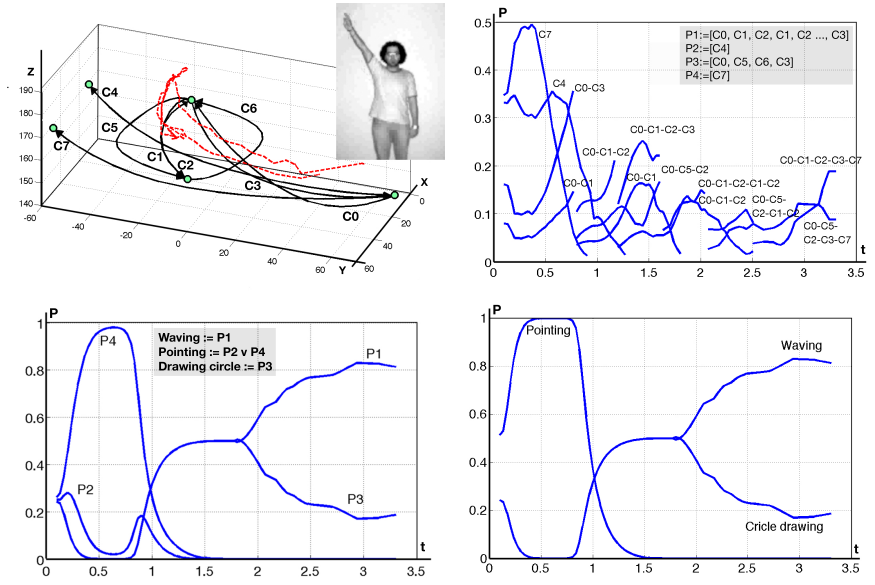

Fig. 3 Simulation results: *top-left:* motor command graph with observed trajectory overlaid (dashed line); *top-right/bottom-left/bottom-right:* changing probabilities of the hypotheses currently entertained on the three motor levels (*commands/programs/schemas*).

At each time point in the observation, one hypothesis corresponds to the most expected movement component. Depending on the number of hypotheses, the maximum expectation value changes over time and the winner threshold is adopted respectively. As shown in Fig. 3 (bottom), the hypotheses first indicate that the observation is similar to a familiar pointing gesture (c_7). Therefore the robot thinks that the user is going to point upwards (p_4). However, after one second the user starts to turn his hand to the right. Thus, the expectation values of the motor commands c_1 and c_5 increase. Consequently, the gestures (p_1 and p_3) attain higher expectancies but the robot still cannot be sure whether the user is going to draw a circle or wave. After about two seconds the movement turns into swinging, which is significantly similar to the waving gestures (p_1) known to the robot. In result, the robot associates the whole movement with the waving schema and can now, e.g., execute a simultaneous imitation using his motor commands in the winning motor program.

8 Conclusion and Outlook

We presented our work towards the establishment of a sensorimotor foundation for social human-robot interaction, guided by neurobiological evidence regarding motor resonance. The model combines hierarchical motor representations with probabilistic forward models and unsupervisedly learned inverse models. Our evaluations with camera data of human gesturing have hitherto produced promising results with respect to a robust recognition and meaningful classification of presented gestures, making use of a growing resonant motor repertoire shared between all sensorimotor processes. The hierarchical nature of the model considers not only the mere spatio-temporal features but also more abstract levels, from the form and trajectory towards the meaning of a gesture. Using a unified motor representation for both perception and action allows direct interactions between these bottom-up and top-down processes and enables the robot to interact in more natural and socially adept ways.

Future work will further extend this line of research and tackle the symmetrical use of the resonant representations for both perceiving and generating gestures, which paves the way toward social human-robot interaction with features like mimicry and alignment. Moreover, since fingers contribute significantly in many co-verbal hand-arm gestures, the introduced finger modules in the model will be realized. Consequently, the setup needs to be extended in order to sense and perceive finger configurations as well. In this context, further training with real gesture data will be necessary to determine the learning capacity of the model.

Acknowledgements This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Center of Excellence in “Cognitive Interaction Technology”.

References

1. R. Amit and M. Mataric. Learning movement sequences from demonstration. In *ICDL '02: Proceedings of the 2nd International Conference on Development and Learning*, pages 203–208, 2002.
2. M. M. Botvinick. Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences*, 12(5):201 – 208, 2008.
3. C. Breazeal, D. Buchsbaum, J. Gray, D. Gatenby, and B. Blumberg. Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots. *Artificial Life*, 11(1-2):31–62, January 2005.
4. S. Calinon and A. Billard. Recognition and reproduction of gestures using a probabilistic framework combining PCA, ICA and HMM. In *22nd International conference on machine learning*, pages 105–112, 2005.
5. K. Dautenhahn. Socially intelligent robots: dimensions of human - robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):679–704, 2007.
6. A. Dijksterhuis and J. Bargh. The perception-behavior expressway: Automatic effects of social perception on social behavior. *Advances in Experimental Social Psychology*, 33:1–40, 2001.
7. C.-L. H. Feng-Sheng Chen, Chih-Ming Fu. Hand gesture recognition using a real-time tracking method and hidden markov models. In *Image and Vision Computing*, volume 21, pages 745–758, 2003.
8. A. Hamilton and S. Grafton. The motor hierarchy: From kinematics to goals and intentions. In *Attention and Performance 22*. Oxford University Press, 2007.
9. M. Haruno, D. M. Wolpert, and M. Kawato. Mosaic model for sensorimotor learning and control. *Neural Computation*, 13(10):2201–2220, 2001.
10. M. Haruno, D. M. Wolpert, and M. Kawato. Hierarchical mosaic for movement generation. *International Congress Series*, 1250:575–590, 2003. Cognition and emotion in the brain. Selected topics of the International Symposium on Limbic and Association Cortical Systems.
11. M. Johnson and Y. Demiris. Hierarchies of coupled inverse and forward models for abstraction in robot action planning, recognition and imitation. *Proceedings of the AISB 2005 Symposium on Imitation in Animals and Artifacts*, 2005.
12. S. Kopp and O. Graeser. Imitation learning and response facilitation in embodied agents. In *Proceedings of the 6th International Conference on Intelligent Virtual Agents*, pages 28–41, Marina Del Rey, CA, 2006.
13. S. Kopp and I. Wachsmuth. Synthesizing multimodal utterances for conversational agents. *Journal of Computer Animation and Virtual Worlds*, 15(1):39–52, 2004.
14. S. Kopp, I. Wachsmuth, J. Bonaiuto, and M. Arbib. Imitation in embodied communication – from monkey mirror neurons to artificial humans. In I. Wachsmuth, M. Lenzen, and G. Knoblich, editors, *Embodied Communication in Humans and Machines*, pages 357–390. Oxford University Press, Oxford, 2008.
15. G. K. Natalie Sebanz. The role of the mirror system in embodied communication. In I. Wachsmuth, M. Lenzen, and G. Knoblich, editors, *Embodied Communication in Humans and Machines*, chapter 7, pages 129–149. Oxford University Press, 2008.
16. S. Schutz-Bosbach and W. Prinz. Perceptual resonance: action-induced modulation of perception. *Journal of Trends in Cognitive Sciences*, 11(8):349–355, 2007.
17. D. M. Wolpert, K. Doya, and M. Kawato. A unifying computational framework for motor control and social interaction. *Philos Trans R Soc Lond B Biol Sci*, 358(1431):593–602, 2003.