In: *Proceedings the 32nd Annual German Conference on Artificial Intelligence, KI 2009*,
pp. 508–515, Paderborn, Germany. doi:10.1007/978-3-642-04617-9_64
The final publication is available at link.springer.com

Requirements and Building Blocks for Sociable Embodied Agents

Stefan Kopp, Kirsten Bergmann, Hendrik Buschmeier, and Amir Sadeghipour

Sociable Agents Group, Cognitive Interaction Technology (CITEC), Bielefeld University
PO-Box 10 01 31, 33501 Bielefeld, Germany

Abstract. To be sociable, embodied interactive agents like virtual characters or humanoid robots need to be able to engage in mutual coordination of behaviors, beliefs, and relationships with their human interlocutors. We argue that this requires them to be capable of flexible multimodal expressiveness, incremental perception of other's behaviors, and the integration and interaction of these models in unified sensorimotor structures. We present work on probabilistic models for these three requirements with a focus on gestural behavior.

1 Introduction

Intelligent agents are nowadays employed as assistants to desktop interfaces, as chatbots on webpages, as instructors in entertainment systems, or as humanoid robots that shall assist household tasks. In in all of these contexts they embody (part of) the user interface with the goal to elevate the interaction between the human and the machine toward levels of natural conversation. However, embodied agents are yet to master a number of capabilities, the most crucial of which are (1) being conversational, i.e., capable of multimodal face-to-face dialogue, (2) being cooperative in reciprocal interaction and joint tasks; (3) being convergent, i.e., able to mutually adapt to and coordinate with a user on a short timescale as well as over longer periods of time, and (4) being companionable, i.e. meet the social dimensions of the former three. All four requirements are interconnected and must be considerd equally important for agents to become sociable. The first one has been tackled in particular in the field of embodied conversational agents (ECAs [1]), the second one in the realm of collaborative systems [2]. In this paper we focus on the third requirement, being convergent.

Natural interaction is characterized by many inter-personal coordinations when individuals feel connected and communicate with ease. For example, behavior congruence, linguistic alignment, interactional synchrony, or fluent back-channeling have been reported (cf. [3]). These mechanisms help to enhance coordination between interacting individuals and significantly eases their joint task of exchanging meaning with signals [4]. We refer to this state as one of "social resonance" to underline the importance of real-time coordination and mutual contingency in the behaviors and mental states of the participants, as well as the dynamics of this interplay. Now, the research question is can we achieve and leverage on such qualities for embodied human-agent interaction? In Sect. 2 we start by analysing which coordinations mechanisms embodied agents would need to be endowed with to that end. We argue that this cannot pertain to a single level

of reciprocating to a particular behavior, but ultimately implies global design criteria for the construction of conversational agents or robots. In Section 3, we focus on three of them with an application to communicative gestures: flexible multimodal expressiveness, incremental understanding through mirroring, and the integration of these models such that the production and the perception of conversational behavior ground and coalesce in the same sensorimotor structures. We present current work that employs state-of-the-art AI techniques to bring these principles to bear in building blocks for interactive agents.

2 Requirement Analysis

Starting from the concept of "social resonance", we note that it embraces many phenomena of face-to-face interaction that have three things in common: (1) they are *interactively contingent*, i.e. their occurrence is causally linked to the interaction context including the partner, (2) they act in a *coordinative* fashion between the interactants, and (3) their occurrence correlates with both the *communicative success*, e.g., fewer misunderstandings, faster goal attainment, less effort in relation to gain, as well as the *social success* of the interaction, e.g., affiliation, prosocial behavior, likelihood of interacting again. The behavioral patterns we are referring to have been described in literature with a lot of of different, often overlapping terms (see [4] for a detailed review). A closer inspection suggests a grouping of different components. In the perspective of conversation as joint action [5] for example, *cooperation and collaboration* are essential for continued dialogic exchange. There, one important coordination device to build a shared basis of common ground, increase familiarity, or lend support is appropriate linguistic and embodied feedback (e.g., backchannels, head-nods). With it, interlocutors continuously show whether they hear and understand each other (or not) and what their attitude towards the speakers current utterance is. Listeners give feedback incrementally (while an utterance is still ongoing) and on different levels of perception and understanding, in order to collaboratively provide closure for single communicative acts and to support the speakers in communicating their thoughts as best as they can – given the situation. Another way of coordinating crops out as *convergence and synchrony* of numerous aspects of the behavior of interaction partners (e.g. lexical choice, phonologic features, duration of pauses, body posture, mimicry). Such phenomena occur fast and lead to behaviors that resemble those of another individuals when we evaluate them positively and when we want to be evaluated positively by them.

Altogether, three kinds of mechanisms can be differentiated, acting on different time scales and serving different coordinative functions: (1) *Behavior coordination* (BHC) lets interactants assimilate their behavior in form, content or timing; (2) *Belief coordination* (BLC) leads to compatible assumptions and convictions, about each other and about specific domains or tasks; (3) *Relationship coordination* (RC) regulates the attitudes and feelings individuals have toward each other. These three kinds of mechanisms bring about inter-personal coordination implicitly and work in parrallel (and jointly) with the commonly conceived exchange of dialog acts. Notably, they are not independent, but connected and inter-related (cf. [4]): BLC is required for RC, as feedback and common ground are prerequisites for establishing familiarity, trust, and rapport. The other way around, a positive relationship (RC) eases belief

coordination and fosters task collaboration. BHC correlates with RC, as mimicry and synchrony are selective and correlate with rapport or affiliation. BHC and BLC are connected, as aligned communicative behavior facilitates person understanding and reflects shared mental representations and common ground. In sum, it is the triad of the coordination mechanisms that creates a state of closeness between interactants that makes the subjective process of constructing meaning-bearing singals better comprehensible and predictable for each other.

Evidence suggests that humans assume, up-front, such qualities also in interactions with artificial interlocutors. It is commonly acknowledged that computers are social actors and this holds in particular for embodied agents, which are known to induce numerous social effects comparable to those when interacting with real humans [6]. Indeed, a growing body of work on embodied agents that can live up to these expectations has started. The most sophisticated technical accounts have been proposed in the realm of "relational agents" [7] and collaborative systems, targeting longer-term qualities like solidarity and companionship. This work pushes standard models of communication by augmenting messages with "social meaning". This allows for deliberate relationship coordination (RC), e.g., by choosing to avoid face threads or to employ social dialogue moves. Others have built agents that recognize or express affective states and show emotional empathy [8, 9], but this is yet to be tied up with the bigger picture of coordinations in dialogic exchange. With regard to short-term behavior coordination agents that mimic a user's head movements are indeed rated more persuasive and positive [10]. Gratch et al. [11] developed a story-listening agent that performs head nods and postural mirroring as rule-based contingent feedback, which was found to increase instant user rapport and to comfort users with social anxiety.

Unlike others [7], we opt for starting with the short-term components of social resonance, i.e., the behavior and knowledge coordinations that facilitate communication right from the start. There is only little modeling work on this topic, and existing systems basically employ simple mapping rules, mainly for the purpose of enabling evaluation studies (e.g. [11]). Like others in social robotics [12], we want to build embodied agents capable of rich social interactions and we explore how we can, to that end, benefit from adopting "design principles" as suggested by recent research on embodied communication [13, 14]. One central principle is to model an unified sensorimotor basis of socio-communicative behavior, and to employ this basis for an incremental behavior perception and understanding, a flexible production of meaningful social actions, and the simulation of coordination mechanisms that are likely to be mediated by this basis. In the following section, we describe approaches that try to bring this principle to bear in work on actual building blocks of sociable embodied agents.

3 Social Behavior Perception-Production Integration

Modeling social resonance requires, on the one hand, to flexibly generate conversational behavior from communicative intentions (top-down). On the other hand, perceiving other agent's social behavior has to be grounded (bottom-up) in the same sensorimotor structures, to connect first-person with third-person knowledge and hypothesize likely interpretations and appropriate responses [14]. We focus on gestural behavior as object of

investigation and briefly present in the following probabilistic models for these processes as well as their fusion.

3.1 Behavior generation with Bayesian Decision Networks

To endow virtual agents with flexible expressiveness we developed a generation model that turns communicative intent into speech and accompanying iconic gestures. Our architecture simulates the interplay between these two modes of expressiveness by interactions between modality-specific modules at each of three stages (Fig. 1, right; for details see [15]): First, *Image Generator* and *Preverbal Message Generator* are concerned with content planning, i.e., they select and activate knowledge from two kinds of knowledge representation. For speech this is a propositional representation of conceptual spatial knowledge. For gesture, we employ Imagistic Description Trees (IDTs; [16]), a computational representation for modeling visuo-spatial imagery of objects shapes. Second, specific planners are integrated to carry out the formation of concrete verbal and gestural behavior. The *Speech Formulator* employs a microplanner using a Lexicalized Tree Adjoining Grammar (LTAG) to generate natural language sentences. The *Gesture Formulator* composes and specifies, on-the-fly, the morphology of a gesture as a typed attribute-value matrix. Finally, *Motor Control* and *Phonation* are concerned with the realization of synchronized speech and gesture animations.

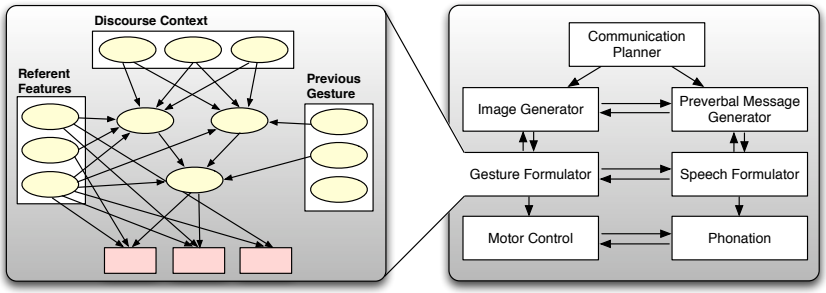


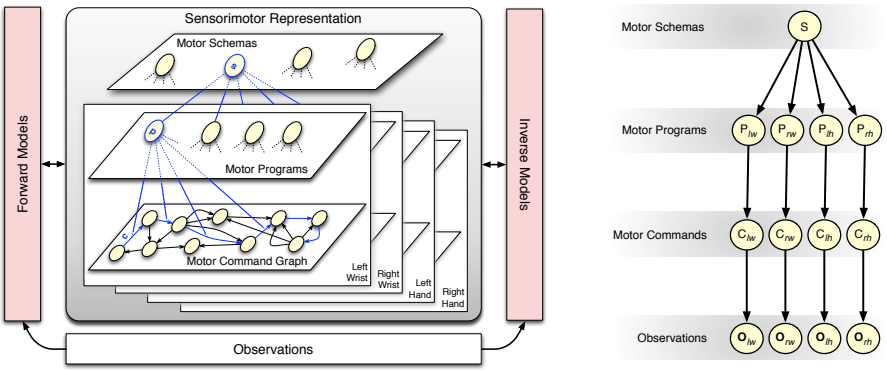
Fig. 1. Overview of the speech and gesture generation model (right), and a zoom in onto the Bayesian decision network for gesture formulation (left).

Especially challenging is the task of gesture formulation as, in contrast to language or other gesture types such as emblems, iconic gestures have no conventionalized form-meaning mapping. Rather, recent findings indicate that this mapping is determined not only by the visuo-spatial features of the referent, but also by the overall discourse context as well as concomitant speech, and its outcome varies considerably across different speakers [15]. As illustrated in the left of Fig. 1, we employ Bayesian decision networks (BDNs) whose structure and probability distributions is learned from empirical corpus data (25 dyads, $\sim$5000 gestures) and then supplemented with decision nodes (red boxes). Influences of three types of variables manifest themselves in dependencies

(edges) between the respective chance nodes: (1) referent features, (2) discourse context, and (3) the previously performed gesture (for details see [17]). BDNs are suitable for gesture formation since they provide a way to combine probabilistic (data-driven) and model-based decision-making. Another rationale of using this method, as discussed shortly, is to prepare the design principle of model integration in order to ground the top-down generation process modeled here in sensorimotor structures that are also the basis of bottom-up gesture perception. Such a model is described next.

3.2 Probabilistic resonances for embodied behavior perception

Embodied behavior perception entails to recruit and actively involve one's own motor structures into the observation of other's behaviors, such that the motor representations that correspond to an observation immediately start to"resonate". We propose a hierarchical sensorimotor system whose layers span from kinematic movement features towards the goal and meaning of a behavior (see Fig. 2(a)). We differentiate between three major levels of a unified sensorimotor representation, modeled as hierarchically connected graphs: (1) The motor command graphs represent motor primitives (controlling small segments of a gestural movement) as edges and the intermediate states as nodes, separately for each body part; (2) a motor program captures the whole movement of a body part performing a gesture and equals a path in the corresponding motor command graph; (3) the motor schema level groups different allowed variants (motor programs) of a gesture into a single cluster. Such a generalization allows to forward the problem of interpreting a gesture from a pure feature analysis to a concurrent, incremental mapping of different aspects of an observation onto own experiences.



(a) (b)

Fig. 2. (a) Hierarchical sensorimotor system for bottom-up grounding in different motor levels for hand-arm gesture perception, (b) the Bayesian network modeling relations between different levels of the sensorimotor representation.

Bayesian inference is applied in utilizing these hierarchical levels for movement perception [18]. A hierarchical Bayesian network (Fig. 2(b)) models the causal influences in-between the different levels. At each level, forward models make probabilistic predictions of the continuation of the movement if it were an instance of a particular motor command, program, or schema, given the evidence at hand and current a priori probabilities. The resulting conditional probabilities refer to the agent's certainty in observing a known movement (see [18] for results when applied to real gesture data). In the case of observing a novel gesture (i.e. probabilities fall below a threshold), inverse models are employed to analyze and then acquire the observed movement into the three representation levels of motor knowledge.

3.3 Towards integrating top-down and bottom-up processing

The integration of the two models for top-down behavior generation and bottom-up perception is subject of ongoing work, confined to an application domain of object and route descriptions. Here we discuss how using a single representational system for both sensory and motor aspects of gestural behavior provides a substrate for perception-action links and, as a result, an architecture for embodied gesture understanding, imitation, and inter-personal coordination of gesture production.

In the latter case, an utterance is planned starting from a communicative goal such as "describe-construction entity-1 subpart-3". As described in Sect. 3.1, the activated parts of the imagistic and propositional content represenations build the basis for speech- and gesture formulation processes (see Fig. 1). The BDN is employed to derive a suitable gesture by specifying a representation technique with certain morphological features, such as handshape, hand orientation or movement trajectory. Now, these values are not handed on to a motor planner, but assigned to the corresponding motor levels in the hierarchical sensorimotor system (see Fig. 2). As a first approach we conceive of assigning these values directly to motor programs in the agent's repertoire (e.g. for drawing a circular trajectory or forming a fist). These activiations percolate probabilistically top-down to the motor command level, which details the planned movement trajectories and postures to be performed. This approach calls for a detailed motor repository of the agent, which becomes mitigated when starting to exploit the motor schema level, e.g., by mapping a general representation technique to a certain motor schema and specifying only the remaining, context-dependent aspects.

This architecture also models an embodied approach to gesture understanding. As described above, perceived movements of the relevant body parts (wrist positions and postures of both hands) activate the most likely motor commands via the forward models. These activations percolate in the Bayesian motor network up to the motor schema level, and only decrease gradually over time. As in the generation case, the *winner* schema and its related motor programs are associated with morphological features which are now associated with the corresponding leaf nodes in the gesture generation BDN. Bayesian networks naturally allow for bi-directional inference, to calculate the likely consequences of causal node states (causal inference) or to diagnose the likely causes of dependent node values (diagnostic inference). Doing the latter the agent can derive likelihoods for a number of contextual parameters that shape gesture use, notably, the visuo-spatial features of the referent object, the discourse context (information structure,

information state), and the communicative goal of the speaker. This inference can be further supported by simply inserting the evidence about the user's previous gesture.

Note that some of the geometrical features of the referent are not encoded in the BDN as chance nodes, but enter gesture generation via rule-based decision nodes. Problematic here is that the use of diagnostic inferences reveals the fact that some nodes feed into both conditional chance nodes and rule-based decision nodes. Consequently, when used in the inverse direction the problem of arbitrating between two value hypothesis arises, one of which is not quantified in terms of certainty. Here, we will employ the form-meaning mapping rules described in [16] to transform morphological features into imagistic representations (IDTs) that will help to disambiguate the determination of the respective features. In the medium term, the solution must be to replace the decision nodes in the generation BDN with chance nodes and to acquire the required larger set of training data in a social learning scenario between the agent and humans. We also note that the resulting imagistic representation will be underspecified in accord with the iconic gesture's vague, evanescent form and characteristically underspecified meaning.

In sum, connecting the generation BDN with the probabilistic sensorimotor structures leads to an agent architecture that grounds social behavior in resonant sensorimotor structures: top-down generation of coverbal gestures results in activations at the motor level, and motor resonances induced by observing a gesture yield its likely interpretations bottom-up. Either way, activation values (probabilities) are not reset directly after perception/generation. Rather we let them decline following a sigmoidal descent towards default values. In this way, the agent's behavior production is biased towards previously perceived behavior (accounting for the so-called *perception-behavior expressway* [19]) and thus allows an embodied agent to engage in gestural alignment and mimicry phenomena between interlocutors (cf. [20]). The other way around, gesture perception is biased towards previously self-generated gestures, which amounts to *perceptual resonance* [21], another suggested mechanism of coordination in social interaction.

4 Conclusion

In this paper, we have analysed which inter-personal coordination mechanisms embodied agents may need to be able to engage in to become sociable. We have pointed out that, as suggested by the sensorimotor grounding of social behavior and intersubjectivity, such agents utlimately need to be based on close integration of models for behavior perception and generation, and an incremental processing on various levels of modularity in a cognitively plausible agent architecture. One stepstone in this direction is the development of a sensorimotor basis in which flexible generation and incremental perception of socio-communicative behavior are grounded in. The work we have outlined here has yielded promising first results and is underway to bring these principles to bear in the development of important building blocks for sociable embodied agents.

Acknowledgements This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in the Collaborative Research Center 673 "Alignment in Communication" and the Center of Excellence in "Cognitive Interaction Technology" (CITEC).

References

1. Cassell, J., Sullivan, J., Prevost, S., Churchill, E., eds.: Embodied Conversational Agents. The MIT Press, Cambridge (MA) (2000)
2. Grosz, B.: Collaborative systems. A.I. Magazine **17**(2) (1994) 67–85
3. Wallbott, H.G.: Congruence, contagion, and motor mimicry: mutualities in nonverbal exchange. In I. Markova, C.F. Graumann, K.F., ed.: Mutualities in Dialogue. Cambridge University Press (1995) 82–98
4. Kopp, S.: From communicators to resonators – coordination through social resonance in face-to-face communication with embodied agents. Speech Communication (accepted)
5. Clark, H.H.: Using Language. Cambridge University Press, Cambridge, UK (1996)
6. Krämer, N.C.: Social effects of virtual assistants. a review of empirical results with regard to communication. In Prendinger, H., Lester, J., Ishizuka, M., eds.: Intelligent Virtual Agents, Springer-Verlag (2008)
7. Bickmore, T.: Relational Agents: Effecting Change through Human-Computer Relationships. PhD thesis, Massachusetts Institute of Technology (2003)
8. Becker, C., Kopp, S., Pfeiffer-Lessmann, N., Wachsmuth, I.: Virtual humans growing up: From primary toward secondary emotions. Künstliche Intelligenz **1** (2008) 23–27
9. Ochs, M., Pelachaud, C., Sadek, D.: An empathic virtual dialog agent to improve human-machine interaction. In: Seventh International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'08). (2008)
10. Bailenson, J.N., Yee, N.: Digital chameleons - automatic assimilation of nonverbal gestures in immersive virtual environments. Psychological Science **16**(10) (2005) 814–819
11. Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R., Morency, L.P.: Virtual rapport. In: Proceedings of 6th International Conference on Intelligent Virtual Agents, Marina del Rey (2006)
12. Breazeal, C., Buchsbaum, D., Gray, J., Gatenby, D., Blumberg, B.: Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots. Artificial Life (2004)
13. Wachsmuth, I., Lenzen, M., Knoblich, G., eds.: Embodied communication in humans and machines. Oxford University Press (2008)
14. Gallese, V., Keysers, C., G., R.: A unifying view of the basis of social cognition. Trends in Cognitive Science **8** (2004) 396–403
15. Bergmann, K., Kopp, S.: Increasing expressiveness for virtual agents–Autonomous generation of speech and gesture. In: Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems. (2009)
16. Sowa, T., Wachsmuth, I.: A model for the representation and processing of shape in coverbal iconic gestures. In: Proc. KogWis05. (2005) 183–188
17. Bergmann, K., Kopp, S.: Gnetic–Using bayesian decision networks for iconic gesture generation. In: Proceedings of the 9th Conference on Intelligent Virtual Agents. (2009)
18. Sadeghipour, A., Kopp, S.: A probabilistic model of motor resonance for embodied gesture perception. In: Proceedings of the 9th Conference on Intelligent Virtual Agents. (2009)
19. Dijksterhuis, A., Bargh, J.: The perception-behavior expressway: Automatic effects of social perception on social behavior. Advances in Experimental Social Psychology **33** (2001) 1–40
20. Kimbara, I.: On gestural mimicry. Gesture **6**(1) (2006) 39–61
21. Schutz-Bosbach, S., Prinz, W.: Perceptual resonance: action-induced modulation of perception. Journal of Trends in Cognitive Sciences **11**(8) (2007) 349–355