Social Resonance and Embodied Coordination in Face-to-Face Conversation with Artificial Interlocutors

Stefan Kopp*

Sociable Agents Group, Cognitive Interaction Technology (CITEC), Bielefeld University, Germany

Abstract

Human natural face-to-face communication is characterized by inter-personal coordination. In this paper, phenomena are analyzed that yield coordination of behaviors, beliefs, and attitudes between interaction partners, which can be tied to a concept of establishing social resonance. It is discussed whether these mechanisms can and should be transferred to conversation with artificial interlocutors like ECAs or humanoid robots. It is argued that one major step in this direction is embodied coordination, mutual adaptations that are mediated by flexible modules for the top-down production and bottom-up perception of expressive conversational behavior that ground in and, crucially, coalesce in the same sensorimotor structures. Work on modeling this for ECAs with a focus on coverbal gestures is presented.

Key words: Social Resonance, Coordination, Embodied Conversational Agents, Gesture

1. Introduction

Computer systems that figure in humanoid form, either as robots or as virtual characters, increasingly meet their users as artificial interlocutors. As so-called Embodied Conversational Agents (ECAs) they embody (part of) the user interface to allow humans to interact with a machine as if having a face-to-face conversation with another human (Cassell et al., 2000b). Such agents can be found nowadays as assistants to desktop interfaces, as chatbots on websites, as tutors in education environments, or as humanoid robots that shall assist in household tasks. They are equipped with abilities for using natural language, conducting dialog, expression emotions, or nonverbal behavior. However, they commonly rely on the classical conduit metaphor of communication (Reddy, 1979): user and agent are taking turns to produce (encode) and receive (decode) meaning-carrying messages that travel across channels between them. As has often been attested (e.g. Scheflen (1982); Gärdenfors (1996)), this "message ping-pong" is an insufficient model of human conversation. Among other things, it cannot account for all those dynamic interactions that interlocutors are often engaged in, to name but a few, linguistic alignment (Pickering and Garrod, 2004), responsive back-channeling (Yngve, 1970), mimicry (Chartrand and Bargh, 1999), or interactional synchrony (Bernieri and Rosenthal, 1991). These phenomena suggest a state of enhanced *coordinations* between interlocutors, which ease the task of exchanging meaning with highy context-dependent messages, and which increase social affiliation and rapport between the interactants (Tickle-Degnen and Rosenthal, 1990; Lakin et al., 2003). Current ECAs, however, are for the most part not capable of such coordinations.

This paper discusses whether these mechanisms can and should be transferred to ECAs, and how this can be achieved. It is structured in two parts. The first part (Sect. 2) analyses relevant social responses and mutual coordinations in human face-to-face conversation. We will sort out mechanisms that affect the interactants' behaviors, beliefs, and attitudes. It will be argued that these mechanisms cannot be treated independently, and that they can be subsumed under the umbrella term *social resonance* (cf. Duncan et al. (2007)). This is to underline the importance of mutual contingency of the interactants' multimodal behaviors, i.e., that the characteristics of the "stimulus" behavior are the same or nearly the same as one's own, as well as the real-time dynamics of this interplay. Resonance also alludes to the mutual "reinforcement" that can be seen to emerge

*Address: Sociable Agents Group, CITEC–Cognitive Interaction Technology, Bielefeld University, P.O. Box 100131, D-33501 Bielefeld, Germany, Tel. +49 521 10612144

Email address: skopp@techfak.uni-bielefeld.de (Stefan Kopp)

as interactants become increasingly coordinated in their interactional characteristics. When being in social resonance, speakers may feel in line and experience their interaction to run smoothly. In this respect, social resonance relates to one component of *rapport*, defined by Tickle-Degnen and Rosenthal (1990) as the feeling of a harmonious or sympathetic connection that interlocutors have when they experience mutual attentiveness, positivity, and coordination. The present paper contributes a systematic analysis of the range and mechanisms of mutual coordination that facilitate conversation as such, and help bringing about rapport in it.

The second part of this paper deals with the role and importance of social resonance for human-agent interaction. In Section 3, we will discuss why the respective coordination mechanisms should be modeled in ECAs and review experiences from existing systems. It will be argued that a major, yet missing step in this direction is to endow agents with an embodied basis for social resonance – flexible models for producing and understanding expressive conversational behavior that ground in and, crucially, are closely coupled through sensorimotor structures. In Section 4 we present work in this direction with a focus on coverbal gestures, and we discuss how this makes ECAs able to engage with their users in coordinations only possible in socially resonant face-to-face communication.

2. Social resonance in human conversation

The concept of social resonance in face-to-face interaction embraces a number of phenomema and aims to point up a principled connection between them. What all these have in common is that (1) they are *interactively contingent*, i.e. their occurrence is directly linked to the interactional context including the partner, (2) they act in a *coordinative* fashion between interactants, and (3) their occurence correlates with the *communicative success*, e.g., fewer misunderstandings, faster goal attainment, less effort in relation to gain, as well as the *social success* of an interaction, e.g., affiliation, prosocial behavior, likelihood of interacting again. We briefly review those phenomena here.

2.1. Mutual coordinations in natural conversation

Coordinative patterns in natural face-to-face conversation have been described in literature with a lot of different, often overlapping terms. In order to disentangle the relevant phenomena, we discriminate between three types that yield coordination of different levels of structure and are likely to operate on different time scales and levels of awareness:

- *Behavior coordination* lets interactants assimilate their behaviors in form, content or timing;
- *Belief coordination* leads to compatible knowledge about specific topics, tasks, or each other;
- *Attitude coordination* regulates the individual's stances toward each other or external objects.

Due to the often vague definitions of the considered phenomena, such an ordering can only be coarse and is not always clear-cut nor unambiguous (Wallbott, 1995). Yet it helps to structure the many intricate mutualities at work in natural conversation, to access their underlying processes, and to inform modeling attempts in ECAs as we shall see later on.

Starting with *behavior coordination* mechanisms, social resonance resides in dialogic exchanges and thus comprises coordinative patterns in linguistic behavior. Giles and Coupland (1991) proposed in their "speech accommodation theory" that individuals try to gain sympathy from their interaction partner by converging, i.e., adapting to each other with respect to their forms or styles of speech. Such convergence has been found, e.g., in dialect, non-verbal behaviors, vocal intensity, prosody, speech rate, phrase duration, or pause length. Pickering and Garrod (2004) noted that speakers in dialog tend to converge on the same expressions, words, phrase structures, etc., and that this allows them to converse with ease and efficiency. They ascribe this "alignment" effect to an automatic "priming" of the interlocutors' lexical, semantic, and syntactic representations, as well as situation models. However, alignment has also been found to be partner-specific (Brennan and Clark, 1996) and thus can be tied also to social functions or explicit audience design efforts (see Branigan et al. (2009) for a recent review).

On a broader note, the term "interpersonal coordination" (Bernieri and Rosenthal, 1991) was used to index behavioral adaptations that occur in natural interaction. Nonverbal aspects of this coordination have also been termed "congruence" (Scheflen, 1964; Kendon, 1973), including identical or similar body postures and synchronized switches of body posture or position. In general, behavioral coordination can be refined into "behavior matching" and "interactional synchrony". Behavior matching refers to the similarity of the behaviors of interaction partners and has been referred to as "congruence", "contingent feedback", "mimicry" (Lakin et al., 2003), or "chameleon effect" (Chartrand and Bargh, 1999). Such a matching has been reported for posture, body movements, facial expressions, mannerisms, verbal complexity, voice loudness, and numerous other

behaviors of the interaction partner. Automatic, nonconscious mimicry has been suggested to act as "social glue" (Chartrand and Bargh, 1999; Lakin et al., 2003) and to build similarity, which 'tends to breed sympathy' Wallbott (1995, p.93). A similar effect has been reported for interactional synchrony (Miles et al., 2009; Bernieri et al., 1994), which refers to the temporal coordination between interactants (also called "entrainment"). Such temporal coordination was found, e.g., in body sway or changes of movement and posture (Condon and Ogston, 1960). Often, this synchronization falls into rhythmic patterns, and it can remarkably be mediated by acoustic signals alone (Shockley et al., 2003).

Turning to the realm of *beliefs*, we first note that communication *per se* is about coordinating the beliefs of interlocutors. We indicate, signal, or display meaning in order for a recipient to understand and share our beliefs and goals, and to respond as intended (Allwood et al., 1992). This belief coordination, in conversation, is a highly dynamic process which has been viewed as demonstrated collaboration or joint action (Clark, 1996): partners coordinate actions in real-time in order to build common ground and achieve a joint task. Mechanisms to enable such cooperative dialog behavior includes "back-channel feedback", by which listeners disclose to speakers their ability and willingness to perceive or understand (Allwood et al., 1992), or demonstrated "grounding" acts by which recipients acknowledge and accept newly presented information (Traum and Allen, 1992). Socially resonant speakers are responsive to such feedback while formulating their next contribution (Brennan and Clark, 1996).

Although belief coordination does not necessarily require behavior coordination, it is revealing that the two often go together. The alignment account by Pickering and Garrod (2004) explicitely states that coordinated behavioral forms precede (by way of priming) coordinated semantic representations. Further, behavior matching can assume a communicative function when it applies to communicative signs. For example, gestural mimicry (Kimbara, 2005), understood as the recurrence of gestural features across speakers, contributes to grounding as it refers by same gestural forms to the same previously established entities. Fig. 1 illustrates one example of gestural mimicry taken from a study on direction-giving dyads (Bergmann and Kopp, 2009b). Here, the recipient demonstrates understanding, i.e., successful belief coordination, by exemplifying it with a different, yet sufficiently similar gesture that mimicks the shape of the original referent.

Finally, there is a level at which interactants try to coordinate their *attitudes* towards each other, as well as the current task goals. For example, speakers who are cooperative will want the others, first, to notice this and, second, to take a similar stance towards their joint project (the conversation itself and maybe also some external task). While a general discussion of this is beyond the scope of this paper, we note that dialog often contains stretches that primarily emphasize social goals; cf. (Bickmore, 2003). "Social dialog frames" can comprise joke-telling, getting acquainted talk, or small talk, and set the stage for attitudinal coordination mechanisms like showing agreement and reciprocal appreciation in order to build solidarity, familiarity, or rapport.

Figure 1: Example of gestural mimicry in natural conversation.

2.2. *Coordination and social resonance*

The previously reviewed phenomena demonstrate that humans in face-to-face communication engage in mutual coordination of (at least) their behaviors, beliefs, and attitudes. The question is: how do those coordinations lead to social resonance between interactants? First of all, it is important to note that coordinations occur frequently and ubiquitously, but not inevitably and not always to the same degree. As Wallbott (1995) put it, dialogical exchange consists of an complex interplay of mutualities and non-mutualities of the interaction partners. Dialog evolves upon a basis of coordinated (grounded) behaviors and beliefs, but is at the same time propelled forward by the nonconformances of the interlocutors. Further, socio-cultural norms, situational demands, or the relationship between two interactants imply boundaries of 'too-much' similarity and mutuality. For example when explicitely detected, high degrees of convergence can backfire and make people feel patronized or uncomfortable Giles (1980). Nevertheless, interlocutors are often found to coordinate themselves with each other more than necessary. Social resonance refers to this very quality of an interaction.

It is also important to note that the different coordination mechanisms are not separate, but go hand in hand (see Fig. 2): belief coordination facilitates attitude coordination as feedback and common ground are prerequisites for establishing familiarity, trust, and rapport.

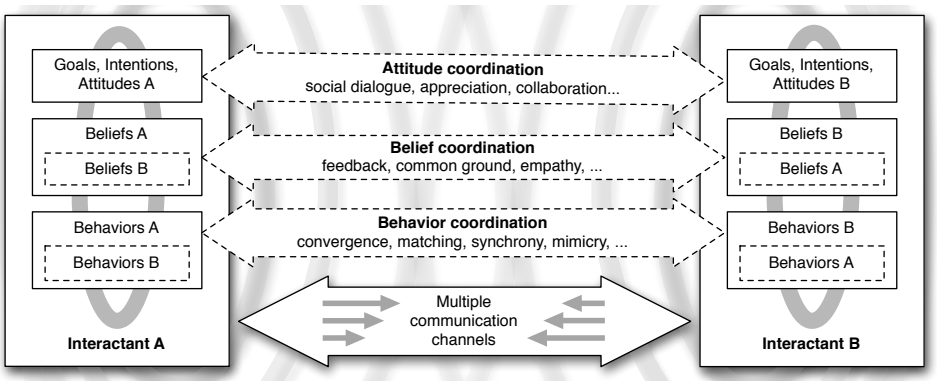


Figure 2: Social resonance exceeds mere information exchange by mutual coordination of behaviors, beliefs, and attitudes.

The other way around, a positive relationship changes belief coordination devices (Cassell et al., 2007) and fosters task collaboration (Bickmore, 2003). Behavior coordination, in turn, is tied to attitude coordination as mimicry and synchrony correlate with affiliation (Lakin et al., 2003), rapport (Miles et al., 2009), and impression management (Uldall et al., 2003). Finally, behavior coordination and belief coordination are linked as aligned communicative behavior reflects shared mental representations and thus common ground (Pickering and Garrod, 2004; Kimbara, 2005). In sum, it is the coordinated interplay of the three kinds of coordination which creates social resonance, each one alone is insufficient. It comes about when speakers are able and willing to be sensitive to and resonate with others on all levels, in support of the interaction and in appreciation of the interlocutor. The resulting state of behavioral and mental alignment, then, is a main ingredient for the feeling of rapport (Tickle-Degnen and Rosenthal, 1990).

One reason that not all interactions nor all speakers are equally amenable to mutual coordinations may lie in differences in the individual's "interpersonal sensitivity" as well as differences in the cooperative and social nature of the activity (Miles et al., 2009; Tickle-Degnen and Rosenthal, 1990). Interpersonal sensitivity (Hall and Bernieri, 2001) is the ability to accurately assess other's abilities, emotional states, and traits from nonverbal cues. Even stronger notions like "empathy" or "emotional contagion" (Hsee et al., 1990) refer to an automatic tendency to mimic another persons emotional experience/expression, and thus to experience/express the same emotions oneself. These capabilities are basic prerequisites for being socially resonant. They are fundamentally rooted in a sensorimotor basis in which fast perception-action links like the human mirror system actively involve the own motor system in the perception of other's actions (Rizzolatti et al., 2001; Wilson and Knoblich, 2005). The resulting "motor resonances" are assumed to underlie behaviors where an individual reproduces, overtly or internally, movements or actions made by another individual. In humans, those resonances have been robustly found for social (not object-directed) behavior (Bertenthal et al., 2006; Montgomery et al., 2007) and this may directly mediate behavior coordinations like mimicry or synchrony. In addition, such processes are involved in empathy and social understanding of others (Gallese et al., 2004), thus reaching into levels of language and social cognition that can mediate the coordination of beliefs and attitudes. Our analysis of the co-occurrence of coordinations, then, suggests that these levels are linked in both directions (bottom-up and top-down) and that these links are particularly effective in socially resonant interactions.

3. Social resonance in human-agent interaction

The previous section analyzed social resonance in face-to-face communication between humans. Now we turn to conversations between humans and embodied artificial systems. Endowing ECAs with capabilities of social resonance has increasingly started to move into focus of researchers. Here we review findings indicating that such a quality is actually relevant for human-agent interaction, and we discuss whether agents need to be coordinating with users themselves. We will thereby also review current approaches to social coordinations in interactive technical systems.

3.1. Why is it relevant?

In human communication, mutual coordinations ease the task of communicating for the interlocutors (com-

municative and cognitive functions) and contribute to socially desirable outcomes of the interaction (social function). Of course, tenants of human-human interactions do not directly translate in human-agent interactions, due to the simplifications and approximations inevitably adopted in agents. However, solid evidence suggests that humans have a propensity to treat agents as social actors and to apply human communication strategies towards them, albeit aware of their machine nature (Reeves and Nass, 1996; Kopp et al., 2005; Krämer, 2008). Correspondingly, humans show behavior coordination effects towards machines. For example, speakers have been reported to align their speech rate, amplitude and pause structure, prosodic contour, or contribution length to computers that speak with a synthetic voice. Alignment is also found in lexical choice as well as syntactic structures (see Branigan et al. (2009) for an excellent overview).

Remarkably, alignment of human users with computers has been found to be similar or even stronger than in human-human communication. Branigan et al. (2009) conclude that alignment with artificial interlocutors is motivated less by social components than by considerations of communicative success, which encompass beliefs about the computer's limited capabilities. The fact, however, that we also find behavior coordination of rather low features, activated by mere exposure to a computer's verbal and nonverbal behavior, speaks to the effectivity of multiple components of coordination (automatic vs. strategic, behaviors vs. beliefs). This effect is likely to be stronger in face-to-face conversation with embodied agents, which are known to induce social effects. Indeed, presenting humans with a sufficiently human-like agent activates our basic sensorimotor mechanisms of social interaction (e.g., Press et al. (2006)) and ECAs may well be able to invoke mimicry or alignment on the part of the human. In sum, social resonance can be an effective phenomenon with ECAs to some extent, and human users readily exhibit different coordination mechanisms towards machines.

3.2. Why building resonant ECAs?

In principle, socially resonant behavior is a hallmark of natural conversation and it seems natural to work towards their inclusion in ECAs. For example, it is obvious that those dynamic mechanisms that help interlocutors coordinate their beliefs (e.g., back-channel feedback or gestural mimicry) can be of great value for conversational artifacts, especially given their limited capabilities, which users cannot readily assess. Further, a growing body of studies indicate that behavior coordinations on the part of the agent may be advantageous: users rate systems more positive, believable, or persuasive when they align with them. This has been reported, e.g., for speech rate, synthesized voices, or head movements (Bailenson et al., 2008). Branigan et al. (2009, p.12) hold that "*just as alignment by a human interlocutor causes people to rate that interlocutor more positively, then, so alignment by a computer induces positive affect towards the computer*". This conforms the role of social resonance in rapport-building also for agents. That is, when we want agents to be able to build rapport with their users, contingent behavior becomes crucial. Gratch et al. (2006) found that a story-listening agent that performs head nods and postural mirroring increases instant user rapport and comforts users with social anxiety (Kang et al., 2008).

Further research in humans has also shown that not being resonant (in the sense of not being imitative and adaptive) can be taken as indicator of social distance and interaction problems, can lower self-esteem in interaction partners (Lakin et al., 2008), or can use up their cognitive resources (Finkel et al., 2006). The latter conforms the hypothesis (Pickering and Garrod, 2004) that behavioral alignment is largely resource-free and even a default for human cognitive processing (due to, e.g., priming effects). Agents that are resonant may well be able to support this cognitive trait. Finally, work on "relational agents" (Cassell and Bickmore, 2001; Stronks et al., 2002) has utilized deliberate attitude coordination, e.g., through avoiding face threads or conducting social dialog. Again, results indicate that this can have positive effects on the interaction with users (Bickmore and Schulman, 2006; Bickmore, 2003).

Overall, we are led to believe that making agents socially resonant can bear several improvements of human-agent interaction. However, whether and in which scenarios exactly human users would value and benefit from resonant agents is an open research question. Research in this direction is needed and important, for it will help our understanding of how HCI works, how robust naturalistic systems can be developed, and what potential limits of human-agent interaction are. Hence we move on to present actual work in this direction. Unlike others, who target long-term coordinations (Cassell and Bickmore, 2001; Bickmore, 2003), we work bottom-up and start with the short-term components of social resonance, i.e., the behavior and belief coordinations that become visible in effects like the gestural mimicry shown in Fig. 1, and which can facilitate communication right from the start. Such low-level mechanisms are likely to rest upon, and partly may even result from, the involvement of sensorimotor components in perceiving, processing, and producing socio-

communicative behavior. For example, the automatic sensorimotor activations discussed in Sect. 2.2 can directly provide a basis (1) for non-consciuous mimicry, when leaking through to motor execution; (2) for understanding the meaning and intentions behind a behavior, when perculating up into higher levels of representing mental states; (3) for alignment, when leaving traces that affect the recruitment of those structures in subsequent behavior production. The next section presents work to enable such embodied coordinations in ECAs.

4. Modeling embodied coordination for ECAs

We start out building socially resonant ECAs by modeling for them embodied coordination of social behavior. This involves three tasks: First, modeling fast and incremental perception of another agent's behavior and its grounding in experiences and action of ones own. Second, building flexible generation models for conversational behavior that are so adjustable and versatile as to enable fine coordination of linguistic and nonverbal behavior. The former runs mainly bottom-up, from concurrently activated sensorimotor structures to hypotheses about intended interpretations and appropriate responses; the latter is to work primarily top-down, driven by communicative intent and mapping into communicative acts in a context-sensitive fashion. Finally, both kinds of processes need to be fused in perception-action structures that can mediate inter-personal coordination. Here we present work that addresses these issues for coverbal gesture.

4.1. Bottom-up gesture perception

Seeing somebody gesturing shall create resonances in the agent's motor system, which then enable the agent to imitate the observed movement overtly or internally. We devised a hierarchical sensorimotor system that spans from kinematic movement features to complex motor structures, and reaches into higher levels of goals and intentions (Sadeghipour and Kopp, 2009). In this view top-down motor control is seen as refinement of commands, from an abstract goal to more detailed motor acts to precise specifications of muscle activations (Hamilton and Grafton, 2008). When perceiving others' social behaviors, activation originates in motor processes and flows bottom-up (Gallese et al., 2004) while being susceptible to top-down modulation, as suggested by the selectivity of mimicry (Lakin et al. 2003).

Our resonance-based gesture perception model is shown in Fig. 3. It is built atop a computational model of gesture motor control for humanoid agents, part of the "Articulated Communicator Engine" (ACE; Kopp and Wachsmuth (2004)). This motor control module provides means of representing a gesture as a composite of spatio-temporal goals, e.g., handshape, orientation of the wrist, or trajectory of the arm movement. These features constitute the lowest level of description of a movement, in terms of the agent's own motor commands. Above this level, we differentiate between three levels of sensorimotor representation: basic motor commands (MC), motor programs (MP), and complex motor schemas (MS); see (Sadeghipour and Kopp, 2009) for detailed descriptions. At the motor command level, movement is represented as a path through a graph, in which nodes denote states of the agent's body and motor system, and edges represent motor commands that cause this system to transition from one configuration into another. Motor programs, at the next level (cf. Fig. 3), are timed paths in the motor command graph. Both structures are present for the left and right hand and arm, respectively. At the top-most level, a motor schema subsumes different motor programs for variants of, say, a waving gesture (e.g., one for waving left at shoulder height and with two large repetitions, and another one for waving right at chest level with three reciprocating movements).

The hierarchical motor structures are the basis on which probabilistic processes of recognizing, imitating, and predicting the behavior of others operate: Visual input of an observed movement is continuously fed into a working memory. Forward models make probabilistic predictions of the possible continuation of the movement if it were a specific motor unit (command, program, or schema). Bayesian evaluation against the actual movement yields conditional probabilities for different motor commands. A Bayesian network models the cross-level activation along probabilistic relationships between the components of different levels. In result, conditional probabilities, interpreted here as degrees of certainty, perculate bottom-up from motor commands to programs and on to schemas. At the same time, activations at higher levels yield predictions that flow top-down and increase/decrease the prior probabilities of lower level candidate structures. In result, as the agent is observing a gesture, resonances come about on all levels in parallel. Fig. 4 illustrates this for an example waving gesture (from motion capture data).

Fig. 5 shows an example from a simulation with two virtual humans, one being the demonstrator of a gesture and the other being the "resonator", whose corresponding own motor program gets activated almost simultaneously and allows imitating the gesture along. This example also demonstrates that the gesture is reproduced

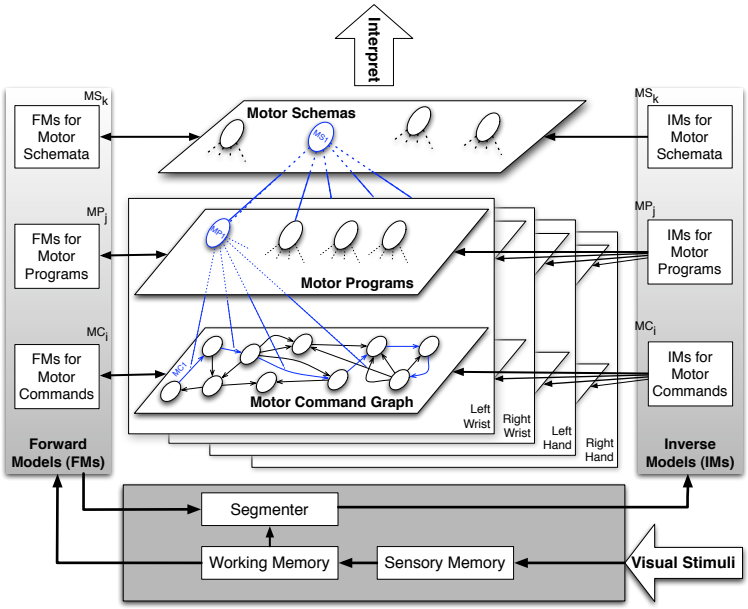


Figure 3: Resonance-based model of bottom-up gesture perception; see text for explanations.

lateralized since, in this case, handedness was considered a significant and potentially meaningful feature (by the motor schema). Thus it needed to be reproduced to demonstrate successful belief coordination, accompanied by an embodied behavior coordination (showing *I know what you mean, I have a similar gestural representation of it, and I know how it feels like*). In this sense gestural mimicry differs from simple mirroring (cf. the example in Fig. 1). This also shows how resonance can arise through mimicry or imitation by indicating shared efficiencies and experiences (Bertenthal et al., 2006). Thus, gestural mimicry has an auxiliary social function in addition to its grounding function. The actually performed contingent gesture, then, is the result of both coordination mechanisms.

Our gesture perception model can simulate for arbitrary social behaviors how motor resonances incrementally emerge, how mimicry comes about, and how activation spreads up into levels of complex, decontextualized motor schemas. But how can this lead to coordination effects in an agent's communicative behavior?

4.2. Top-down speech and gesture generation

The second cornerstone of socially resonant ECAs is highly flexible behavior generation. Gestures have been shown to vary in size, redundancy with speech, or complexity, depending on the grounding status of the information encoded or the meaning-form pairings employed (Kimbara, 2005). When looking at continued interaction in which social resonance should emerge, gestures play an important role as they indicate engagement, draw attention, and convey information about the speaker's mental state not present in speech. This holds especially for iconic gestures that create depictions of objects or events in space (McNeill, 1992), thus accompanying verbal descriptions with imagistic information. Gestural coordination thus can have many possible functions, from demonstrating engagement to building positive affect, to confirming successful coordination of perceptual mental states and their grounding in shared sensorimotor experiences.

Only a full-blown production model that comprises rich multimodal knowledge structures, embedded in discourse, and can turn selected parts of them into coordinated words and gestures would provide the flexibility needed for all these coordinations. We thus devised a generation model, outlined in Fig. 6(a), that implements the main stages of deriving words and gestures, as well as the presumed interactions between the two modes of expressiveness along these stages (Bergmann and Kopp, 2009b). The major steps are content planning (i.e., figuring out what to convey), behavior formulation (i.e., figuring out how to convey it best), and realization (i.e., conveying it). Each step is realized in a modality-specific way: in the speech branch, a Message Generator draws upon a propositional knowledge representa-

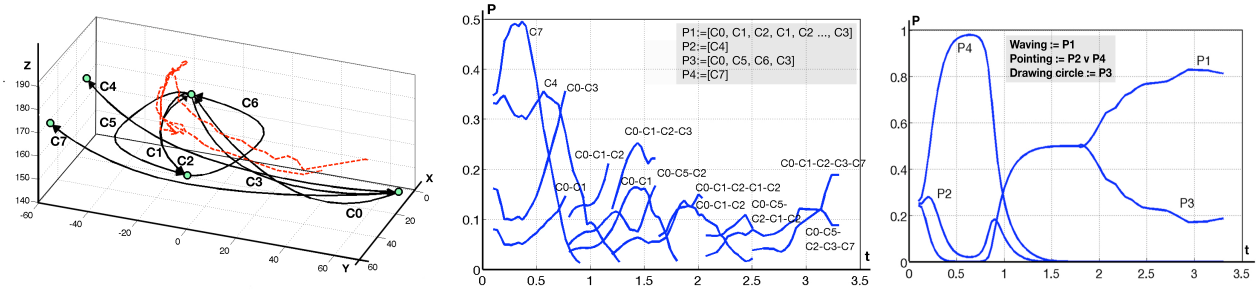


Figure 4: Example of probablistic gesture perception of a waving gesture (dotted; *left*) with the evolving probabilities of the agent's own corresponding motor commands (*middle*) and motor programs (*right*). The motor program **P1** is eventually singled out already during observation.

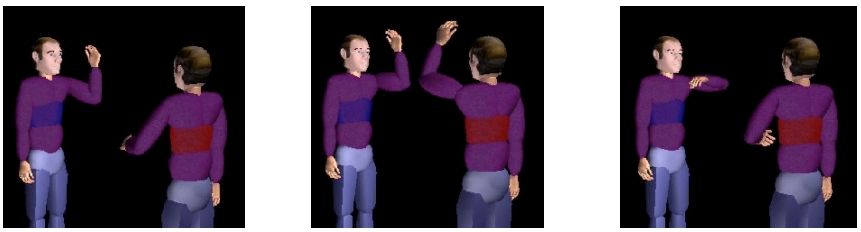


Figure 5: *Left*: one interlocutor (left) performs a simple hand-raising gesture, the observer (right) has immediate motor resonances and imitates simultanesously; *middle*: motor-level intentions were successfully reproduced; *right*: demonstration continues but the movement is new to the observer who returns to a rest position for learning.

tion and selects facts that a Speech Formulator turns into linguistic utterances, which are then synthesized using our realization engine ACE (Kopp and Wachsmuth, 2004). For gesture, an Image Generator operates upon a simulated visuo-spatial imagery (Imagistic Description Trees (IDT); Sowa and Wachsmuth (2005)) and realizes, e.g., spatial perspective taking. Speech and gesture generation interact at the level of semantic representation via multimodal concepts, creating bindings of IDTs with corresponding propositional formulations.

In recent work, we have demonstrated that this flexible generation model can be used to simulate verbal lexical alignment (Buschmeier et al., 2009). Of particular importance here is the Gesture Formulator, which is to find gestural depictions of selected parts of visuo-spatial imagery. To this end, it is not restricted to a fixed set of self-contained gestures, but constructs them on the fly from single morphological features (handshape, wrist location, palm direction, extended finger direction, movement trajectory and direction) using a Bayesian decision network as shown in Fig. 7 (Bergmann and Kopp, 2009a). The network was learned from a large corpus of speech and gesture behavior in direction-giving (25 dyads; about 5000 gestures). In these data we found situational factors like referent shape or speech

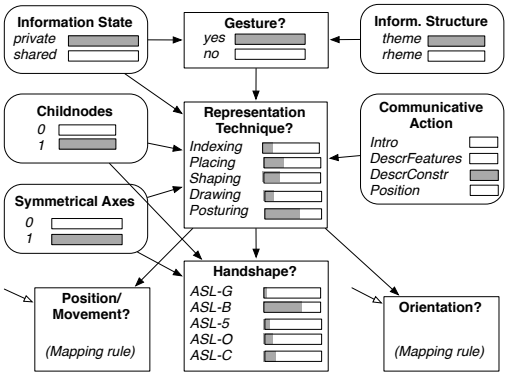
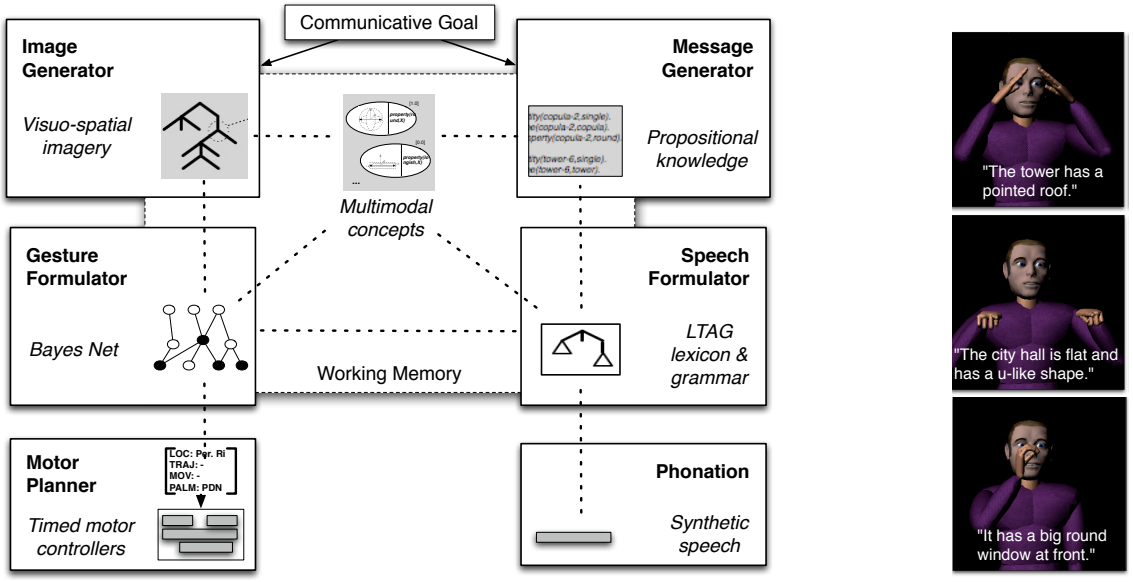


Figure 7: Bayesian decision network for gesture formulation.

act type to have a systematic impact on the overall gesture representation technique employed (shaping, drawing, posturing, placing or pointing). On the other hand, there are significant individual differences in whether people make a gesture at all or which gesture features they prefer. The decision network allows us to model these probabilistic dependencies learnt from speaker data, and to combine them in a unified manner with



(a) (b)

Figure 6: (a) Schematic of the generation model for speech and iconic gesture; (b) exampe utterances produced with it.

decision nodes for features that must be derived in a model-based way to ensure iconicity of the gesture (e.g., position and orientation). Fig. 6(b) shows example utterances produced with the virtual human Max, autonomously generated by our model from simple communicative goals like "*describe churchtower-1 roof-3*".

4.3. Fusing perception and generation

The final step is to fuse the previous two models such that perception-induced motor resonances can affect behavior production. This is part of an ongoing research program and, here, we discuss the straightforward approach to this, namely, to ground gesture production in the sensorimotor structures of the perception model. This means to connect the productive rule-based Gesture Formulator, which was built to compose new gestures from scratch, with a more exemplar-based approach that rests upon previously experienced and learnt own motor patterns. More precisely, we propose to connect the Bayesian network that models the probabilistic dependencies and choices in deciding gesture form, with the probabilistic network of motor schemas and motor programs as follows.

Gesture generation utilizes a network that associates pragmatic and semantic aspects with features of a gesture like representation technique, handedness, hand shape or position, and possibly also schematized feature combinations (e.g., moving along a certain trajectory while holding the hand orthogonal to the direction of movement). The according decision nodes, now, need to connect to motor structures and formulate probabilities over these value sets, and the levels of motor schemas and motor programs seem to be well suited for this: motor schemas can be associated with complete gestures or iconic representation techniques, motor programs with certain morphological features. That way, motor structures also come to be endowed with a "semantic potential". The model-based decision nodes can be extended to work bi-directionally, which we reckon to be possible since inverted forms of the mapping rules have been applied in earlier work on gesture interpretation (Sowa and Wachsmuth, 2005). Finally, the network can easily be made dynamical be altering probabilities temporarily and let them slowly return to long-term distributions.

This approach has several advantages. First, Bayesian networks allow for both predictive and diagnostic inferences. The combined model can thus be used for incrementally inferring the likely interpretations of a gesture or, more precisely, the motor resonances it evokes. Interpretations thereby automatically arrive at all variables modeled in the network, including imagistic meaning, information structure and speech act type, or the spatial perspective taken. Second, the network dynamics allows for modeling preactivations over feature values, which is needed to simulate alignment. This is easily achieved by adjusting a priori distributions and incorporating them in the Bayesian diagnostic and predictive inferences, as well as the de-

cision nodes. Finally, the resultant network structure will connect meaning-based structures with form-based structures, and it will thus allow seamless resonance-spreading across them. This is crucial as we find in our gesture data that behavior coordination and belief coordination intermingle (as in the example in Fig. 1). Finally, in an integrated model as proposed here, in which imagistic meaning interfaces with language semantics, this can even extend into supramodal effects, e.g., when perceiving a gesture lets the agent use different, semantically coordinated language.

4.4. Example

To illustrate how the fused model would work, we give an example of how an agent can dynamically coordinate behavior and beliefs with a user. The simplest case is that the agent, in conversation with a human user, receives a user's utterance like "please give me the plate" including a static iconic gesture adding to speech complementary information about the plate's shape. As described above, bottom-up gesture perception directly leads to fast probability distributions in the Bayesian gesture generation network, as to the likely interpretation of this gesture (e.g., that it refers to a round-like referent that is new to the discourse). This allows, first, for fast affirmative feedback (e.g. head nod) and, second, for further reasoning that in this case might result, e.g., in a communicative goal "*confirm give plate-2*". From this, the generation model produces an utterance "ok the round one", accompanied by a gesture that is, thanks to shifted probability distributions, similar to the user's gesture, i.e., static modeling as opposed to dynamic circle-drawing, which the model could have come up with as well. Note that a different gesture by the user, or the absence of it, would have caused the agent automatically to behave differently. In result the user experiences an agent that is engaged and responsive (gives fast feedback), shares the same gesture motor repertoire, is collaborative (demonstrates this personal common ground), and aligns (gesture mimicry).

5. Discussion and Conclusion

Natural face-to-face conversation is characterized by qualities beyond the discrete exchange of clear-cut messages. A theoretical analysis has revealed numerous ways in which interlocutors coordinate their behaviors, beliefs, and attitudes. These mechanisms serve communicative, cognitive, and social functions, and this points us to the fact that we need to pay more attention to socio-communicative factors in ECA design. It is not claimed here that human-agent communication can or should be made identical to human-human encounters, especially with regard to social aspects. Neither it is claimed that agents can engage in full-blown social relationships with humans. What we argue, however, is that mechanisms like mimicry, alignment, and synchrony are essential coordination devices in face-to-face conversation, and that it may be a significant improvement of human-agent interaction to also impart those mechanisms to embodied conversational artifacts. Enabling such mechanisms for human-machine interaction may probably be one asset of ECA-based interfaces.

Endowing machines, agents or robots, with better capabilities for social interaction has been a goal of researchers for quite some time. Current virtual agents, as partly discussed in Sect. 3, have focused either on establishing and maintaining relationships with users over time, or have looked at instant rapport. The latter systems share with the present work a focus on short-term contingencies and coordination. However, no work has tried so far to develop deep generation and perception models for conversational speech-gesture behavior, and to integrate them based on tenets of embodied coordination. Previous approaches to iconic gesture generation, e.g. (Cassell et al., 2000a; Kopp et al., 2004), employed predefined lexicons and do not explicate gesture semantics, such that they lack sufficient flexibility to achieve viable behavior and belief coordination.

Likewise, classical approaches to gesture perception rely on pattern classification. In recent times, a growing body of work in social robotics is directed to robots capable of rich social interactions with humans. This includes movement perception for imitation based on partially shared sensory and motor representations (Amit and Mataric, 2002; Shon et al., 2007). Billard (2002) employed imitation in a robot in order to create a shared perceptual context for learning signs. (Breazeal et al., 2005) have proposed a full simulation-theoretic architecture with integrated perception-action structures, based on which a robot acquires social behavior and then infers 'empathetic' affective states or beliefs of the human with whom it interacts. The work presented here is inspired by this approach and exceeds it in two respects: first, we target the perception and generation of complex communicative signs including speech and iconic gesture; second, the perception model presented here allows for incremental processing and fast, contingent behavior. Work like Breazeal et al. (2005)'s point the way toward using imitation and empathy as a means of social learning, grounding behavior perception and action in an actual physical body, and boot-strapping representational aspect of a Theory of

Mind that captures commonalities as well as differences between an agent's own states and the states of others. But what is missing yet, are sufficiently powerful models of communicative robot behavior, such that a robot could actually engage with a human user in coordinating beliefs and behaviors at eye level. Our work explores this direction in the realm of speech and gesture for virtual agents. Our models have provided very promising results so far, and we are confident that their integration as presented and discussed here will elevate the interaction abilities of artificial interlocutors to a new level, at which contingent behavior is not coincidental but an earmark of socially resonant human-agent interactions. This will then enable first opportunities for thorough empirical testing of social resonance with ECAs, and the possibilities and limits of human-agent interaction more generally.

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

References

Allwood, J., Nivre, J., Ahlsén, E., 1992. On the semantics and pragmatics of linguistic feedback. Journal of Semantics 9, 1–26.

Amit, R., Mataric, M., 2002. Learning movement sequences from demonstration. In: Proc. 2nd Int. Conf. on Development and Learning (ICDL'02). pp. 203–208.

Bailenson, J. N., Yee, N., Patel, K., Beall, A. C., 2008. Detecting digital chameleons. Computers in Human Behavior 24, 66–87.

Bergmann, K., Kopp, S., 2009a. Gnetic - using bayesian decision networks for iconic gesture generation. In: Intelligent Virtual Agents. LNAI 5773. Springer-Verlag, pp. 76–89.

Bergmann, K., Kopp, S., 2009b. Increasing expressiveness for virtual agents–autonomous generation of speech and gesture. In: Proceedings of AAMAS 2009. pp. 361–368.

Bernieri, F., Rosenthal, R., 1991. Coordinated movement in human interaction. In: Feldman, Rime (Eds.), Fundamentals of nonverbal behavior. Cambridge University Press, New York, pp. 401–432.

Bernieri, F. J., Davis, J. M., Rosenthal, R., Knee, C. R., 1994. Interactional synchrony and rapport: Measuring synchrony in displays devoid of sound and facial affect. Personality and Social Psychology Bulletin 20 (3), 303–311.

Bertenthal, B. I., Longo, M. R., Kosobud, A., 2006. Imitative response tendencies following observation of intransitive actions. Journal of Experimental Psychology: Human Perception and Performance 32 (2), 210–225.

Bickmore, T., 2003. Relational agents: Effecting change through human-computer relationships. Ph.D. thesis, Massachusetts Institute of Technology.

Bickmore, T., Schulman, D., 2006. The comforting presence of relational agents. In: Proceedings of CHI 2006. pp. 550–555.

Billard, A., 2002. Imitation: a means to enhance learning of a synthetic proto-language in an autonomous robot. In: Dautenhahn, K., Nehaniv, C. (Eds.), Imitation in animals and artifacts. MIT Press, Cambridge, MA, pp. 281–310.

Branigan, H., Pickering, M., Pearson, J., McLean, J., 2009. Linguistic alignment between humans and computers, (in press).

Breazeal, C., Buchsbaum, D., Gray, J., Gatenby, D., Blumberg, B., 2005. Learning from and about others: toward using imitation to bootstrap the social understanding of others by robots. Artificial Life 11 (1-2), 31–62.

Brennan, S., Clark, H., 1996. Lexical choice and conceptual pacts in conversation. Journal of Experimental Psychology: Learning, Memory And Cognition 22 (6), 1482–93.

Buschmeier, H., Bergmann, K., Kopp, S., 2009. An alignment-capable microplanner for natural language generation. In: 12th Eur. Workshop on Natural Language Generation. pp. 82–89.

Cassell, J., Bickmore, T., 2001. A relational agent: A model and implementation of building user trust. In: Proceedings of CHI'01. pp. 396–403.

Cassell, J., Gill, A., Tepper, P., 2007. Coordination in conversation and rapport. In: ACL Workshop on Embodied Language Processing. pp. 40–50.

Cassell, J., Stone, M., Yan, H., 2000a. Coordination and context-dependence in the generation of embodied conversation. In: Proceedings of the First International Conference on Natural Language Generation. pp. 171–178.

Cassell, J., Sullivan, J., Prevost, S., Churchill, E. (Eds.), 2000b. Embodied Conversational Agents. MIT Press, Cambridge, MA.

Chartrand, T., Bargh, J., 1999. The chameleon effect: The perception-behavior link and social interaction. Journal of Personality and Social Psychology 76, 893–910.

Clark, H. H., 1996. Using Language. Cambridge University Press.

Condon, W. S., Ogston, W. D., 1960. Sound-film analysis of normal and pathological behavior patterns. Journal of Nervous and Mental Disease 143, 338–347.

Duncan, S., Franklin, A., Parrill, F., Welji, H., 2007. Cognitive processing effects of 'social resonance' in interaction. In: Proceedings Gesture 2007 - the conference of the Int. Society of Gesture Studies. Evanston, IL.

Finkel, E., Campbell, W., Brunell, A., Dalton, A., Scarbeck, S., Chartrand, T., 2006. High-maintenance interaction: inefficient social coordination impairs self-regulation. J Pers Soc Psychol 91 (3), 456–475.

Gallese, V., Keysers, C., G., R., 2004. A unifying view of the basis of social cognition. Trends in Cognitive Science 8, 396–403.

Gärdenfors, P., 1996. Human communication: What happens? In: Reichert, B. (Ed.), The Contribution of Science and Technology to the Development of Human Society.

Giles, H., 1980. Accomodation theory: Some new directions. In: DeSilva, S. (Ed.), Aspects of Linguistic Behaviour: A Festschrift in Honour of Robert Le Page. University of York Press, pp. 105–136.

Giles, H., Coupland, N., 1991. Language: Contexts and Consequences. Wadsworth Publishing, Belmond, CA.

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R., Morency, P.-L., 2006. Virtual rapport. In: Proceedings of IVA '06. LNAI 4133. Springer-Verlag, pp. 14–27.

Hall, J., Bernieri, F. (Eds.), 2001. Interpersonal Sensitivity. Lawrence Erlbaum Assoc., Mahwah, NJ.

Hamilton, A., Grafton, S. T., 2008. The motor hierarchy: from kinematics to goals and intentions. In: Rosetti, Y., Kawato, M., Haggard, P. (Eds.), Attention and Performance. Vol. 22. Oxford University Press.

Hsee, C. K., Hatfield, E., Carlson, J., Chemtob, C., 1990. The effect of power on susceptibility to emotional contagion. Cognition and Emotion 4, 327–340.

Kang, S.-H., Gratch, J., Wang, N., Watt, J., 2008. Does contingency of agents' nonverbal feedback affect users' social anxiety? In: Proceedings of International Conference on Autonomous Agents and Multiagent Systems (AAMAS '08). pp. 120–127.

Kendon, A., 1973. The role of visible behavior in the organization of social interaction. In: Cranach, M., Vine, I. (Eds.), Social communication and movement. Wiley, London, pp. 29–74.

Kimbara, I., 2005. On gestural mimicry. Gesture 6 (1), 39–61.

Kopp, S., Gesellensetter, L., Krämer, N., Wachsmuth, I., 2005. A conversational agent as museum guide - design and evaluation of a real-world application. In: Intelligent Virtual Agents. LNAI 3661. Springer-Verlag, pp. 329–345.

Kopp, S., Tepper, P., Cassell, J., 2004. Towards integrated microplanning of language and iconic gesture for multimodal output. In: Proceedings of the International Conference on Multimodal Interfaces (ICMI'04). pp. 97–104.

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

Krämer, N. C., 2008. 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, pp. 507–508.

Lakin, J., Chartrand, T., Arkin, R., 2008. I am too just like you: nonconscious mimicry as an automatic behavioral response to social exclusion. Psychol Sci 19 (8), 816–822.

Lakin, J., Jefferis, V., Cheng, C., Chartrand, T., 2003. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. Nonverbal Behavior 27 (3), 145–162.

McNeill, D., 1992. Hand and Mind - What Gestures Reveal about Thought. University of Chicago Press, Chicago.

Miles, L., Nind, L., Macrae, C., 2009. The rhythm of rapport: interpersonal synchrony and social perception. Journal of Experimental Social Psychology, 585–589.

Montgomery, K., Isenberg, N., Haxby, J., 2007. Communicative hand gestures and object-directed hand movements activated the mirror neuron system. Social cognitive and affective neuroscience 2, 114–122.

Pickering, M. J., Garrod, S., 2004. Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences 27, 169–226.

Press, C., Gillmeister, H., Heyes, C., 2006. Bottom-up, not top-down, modulation of imitation by human and robotic models. European Journal of Neuroscience 24, 2415–2419.

Reddy, M., 1979. The conduit metaphor – a case of frame conflict in our language about language. In: Ortony, A. (Ed.), Metaphor and thought. Cambridge University Press, Cambridge, UK, pp. 284–297.

Reeves, B., Nass, C. I., 1996. The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press.

Rizzolatti, G., Fogassi, L., Gallese, V., 2001. Neurophysiological mechanisms underlying the understanding and imitation of action. Nature Reviews Neuroscience 2, 661–670.

Sadeghipour, A., Kopp, S., 2009. A probabilistic model of motor resonance for embodied gesture perception. In: Intelligent Virtual Agents. LNAI 5773. Springer-Verlag, pp. 80–103.

Scheflen, A., 1982. Comments on the significance of interaction rhythms. In: Davis, M. (Ed.), Interaction Rhythms. Human Sciences Press, pp. 13–22.

Scheflen, A. E., 1964. The significance of posture in communication systems. Psychiatry 27, 316–321.

Shockley, K., Santana, M., Fowler, C. A., 2003. Mutual interpersonal postural constraints are involved in cooperative conversation. Journal of Experimental Psychology: Human Perception and Performance 29, 326–332.

Shon, A., Storz, J., Rao, R., 2007. Towards a real-time bayesian imitation system for a humanoid robot. In: 2007 IEEE Int. Conf. on Robotics and Automation. pp. 2847–2852.

Sowa, T., Wachsmuth, I., 2005. A model for the representation and processing of shape in coverbal iconic gestures. In: Proc. KogWis05. Schwabe, Basel, pp. 183–188.

Stronks, B., Nijholt, A., van der Vet, P., Heylen, D., 2002. Friendship relations with embodied conversational agents: Integrating social psychology in eca design. In: Proceedings CHI '02 Workshop Philosophy and Design of Socially Adept Technologies. pp. 25–28.

Tickle-Degnen, L., Rosenthal, R., 1990. The nature of rapport and its nonverbal correlates. Psychological Inquiry 1 (4), 285–293.

Traum, D., Allen, J., 1992. A "speech acts" approach to grounding in conversation. In: Second International Conference on Spoken Language Processing (ICSLP'92). pp. 137–140.

Uldall, B., Hall, C., Chartrand, T. L., 2003. Optimal distinctiveness theory and mimicry: When being distinct leads to an affiliation goal and greater nonconscious mimicry, manuscript in preparation, Ohio State University.

Wallbott, H. G., 1995. Congruence, contagion, and motor mimicry: mutualities in nonverbal exchange. In: I. Markova, C.F. Graumann, K. F. (Ed.), Mutualities in Dialogue. Cambridge University Press, Ch. 4, pp. 82–98.

Wilson, M., Knoblich, G., 2005. The case for motor involvement in perceiving conspecifics. Psychological Bulletin 131 (3), 460–473.

Yngve, V., 1970. On getting a word in edgewise. In: Papers from the 6th Regional Meeting of the Chicago Linguistics Society. University of Chicago, pp. 567–578.