Giving Interaction a Hand — Deep Models of Co-speech Gesture in Multimodal Systems

Stefan Kopp
Faculty of Technology / CITEC
Bielefeld University
+49 521 106 12144
skopp@techfak.uni-bielefeld.de

ABSTRACT
Humans frequently join words and gestures for multimodal communication. Such natural co-speech gesturing goes far beyond what can be currently processed by gesture-based interfaces and especially its coordination with speech still poses open challenges for basic research and multimodal interfaces alike. How can we develop computational models for processing and generating natural speech-gesture behavior in a flexible, fast and adaptive manner similar to humans? In this talk I will review approaches and methods applied to this problem and I will argue that such models need to (and can) be based on a deeper understanding of what shapes co-speech gesturing in a particular situation. I will present work that connects empirical analyses with computational modeling and evaluation to unravel the cognitive, embodied and socio-interactional mechanisms underlying the use of speech-accompanying gestural behavior, and to develop deeper models of these mechanisms for interactive systems such as virtual characters, humanoid robots, or multimodal interfaces.

Categories and Subject Descriptors
H.1.2 [User/Machine Systems]; H.5.2 [User Interfaces]: interaction styles, natural language; I.2 [Artificial Intelligence]: cognitive simulation; I.2.7 [Natural Language Processing]: language generation.

General Terms: Algorithms, Experimentation

Keywords
Gesture; speech; multimodal systems; generation; recognition

1. INTRODUCTION
Gesture-based interaction is experiencing a renaissance. The renewed interest is fueled by the availability of unobtrusive and affordable sensing technology, as well as the need for new interaction paradigms for devices like Smart TVs or touchpads. The usual approaches build on bodily movements that can be reliably detected and that map onto physical manipulation, e.g. flick or pinch. This, however, is still far from what humans intuitively do with their hands when engaged in a multimodal interaction -- co-speech gesturing.

Natural co-speech gesturing encompasses a variety of visible actions (with the hands or with other body parts) used as an utterance, or as part of an utterance (Kendon 2004). I will focus on hand gestures performed while speaking.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s). Copyright is held by the author/owner(s).

ICMI'13, December 9–13, 2013, Sydney, Australia.
ACM 978-1-4503-2129-7/13/12
http://dx.doi.org/10.1145/2522648.2532201

They are frequent parts of multimodal deliveries and they add a great deal to the flexibility, efficiency and resilience of multimodal communication. Speakers move their hands in many different ways and in service of multiple functions (cf. McNeill 2005), for instance to pace and mark focal elements in their speech (temporal highlighting) and to depict visuospatial aspects of the object or action being described (iconicity). In that way different kinds of multimodal ensembles are orchestrated, in which a gesture may complement speech (e.g., "it is this big" + size-indicating gesture), supplement it (e.g., "there’s window" + shape-depicting gesture) or convey mostly redundant information (e.g., "there’s a round window" + shape-depicting gesture). The last decades of gesture research have shown many factors to influence the specific multimodal behavior a speaker exhibits (from communicative intent to language resources and skills, cognitive state, individual characteristics, or cultural background; e.g. see Alibali 2005). That is, how speech and gestures are used together is highly revealing, and allowing unrestrained co-speech gesturing can potentially be very useful for multimodal systems.

Unfortunately, after a great interest in speech-gesture interfaces in the 90’s, it seems that natural gesturing has been deemed too hard for automatic processing systems. Another caveat is that gesture processing has been treated detached from approaches to generating speech and gesture in embodied agents, which has received continued research up to present. In both realms, however, the complexity of the phenomenon is not fully understood and nowadays many simplifying (often misleading) assumptions or rules of thumb are used, to the result that technical implementations are rigid and brittle.

In this talk I will present research on models for processing and generating natural speech-gesture behavior, in a flexible, fast and adaptive manner that is similar to humans. Using empirical analyses as well as computational modeling and evaluation, I will argue that such models need to (and can) be based on a more complete picture of what shapes co-speech gesturing in a particular situation. I will in particular present work on integrated deep models of the cognitive, embodied and socio-interactional mechanisms underlying the use of speech-accompanying gestural behavior in multimodal interaction.

1.1 Cognitive production models
Despite a long-standing research work in Psychology/Psycho-linguistics on language and gesture production, there is still no consensus as to the detailed underlying cognitive processes. We have developed a larger production architecture that integrates the planning and coordination of content, form, and behavior in both modalities (Bergmann & Kopp 2009). It is built atop a computational model of gesture motor control that provides means of forming a gesture as a composite of spatiotemporal goals (e.g., hand-shape, orientation of the wrist, or trajectory of the arm...
movement) and that has been used to drive multimodal behavior of virtual conversational agents and humanoid robots.

The architecture comprises a cognitive production model that demonstrates how the coordination of words and (iconic) gestures can be grounded in conceptualization processes that operate upon a dynamically shaping multimodal memory (Kopp et al. 2013). The model predicts, e.g., how linguistic or cognitive constraints affect overt multimodal delivery (reproducing empirical findings; Bergmann et al. 2013). Going beyond simple rules of thumb like the often over-simplified temporal synchrony rule, such a model paves the way for a more detailed account, e.g., of how the degree of multimodal coordination we see between words and gestures can provide information about the speaker’s linguistic skills or cognitive state. This also suggests a way for detailed accounts of how personal style and individual differences manifest themselves in gestural behavior—an essential step to leap forward to robust recognition and processing systems.

1.3 Embodied processing principles

Seeing somebody gesturing and understanding the intended meaning is not a purely analytical inference process. Drawing inspiration from the involvement of motor brain areas in action perception, we assume that gesture perception, too, can be framed as continuous evaluation of internal predictions to form and test hypotheses based on one’s own motor expertise. I will present different probabilistic models for incremental processing of visual input on gesture trajectories. Shallow models allow for fast recognition and classification; making them ‘deeper’ in the sense of adding higher layers of a sensorimotor hierarchy enables concurrent interpretation. I will present a model that spans from kinematic movement features (in terms of the agent’s own motor commands), to abstract gesture schemas reaching into higher levels of goals and intentions (Sadeghipour & Kopp, 2011). By way of probabilistic prediction of possible continuations of a movement, and Bayesian evaluation against the actual movement, probabilistic activations of motor structures emerge during observation. A Bayesian network allows this activation to spread across levels, such that probabilities percolate bottom-up from motor commands to schemas, while predictions and previous posteriors flow top-down as prior probabilities. In result, as the agent is observing a gesture, resonances come about on all levels in parallel.

This work also demonstrates how to provide structures that support both perception and generation. The fact that both processes, at least partially, operate upon the same representational structures accounts naturally for interpersonal alignment effects, when perception-induced activations as increased priors affect the recruitment of motor structures during subsequent behavior production. The model also provides grounds to explore how social behaviors that are short-lived and created flexibly on the spot, such as co-speech gestures, can be learned and acquired by technical systems. I will discuss work that investigates how to learn schematic, decontextualized representations of a gesture (e.g., waving) that separate variant features (e.g., handedness) from its invariant characteristics (e.g., hand shape).

1.2 Social and interactional nature

The main difference between movements used in gesture-based interfaces and human gestures in natural dialogue, is that the latter are not fixed but dynamically evolving. Speech and gesture together are contingent upon the dynamically co-constructed dialogue context and the continuous coordinations taking place between the interlocutors. Examples of such effects include the grounding of novel gestural forms, the reduction of gestures over repeated uses, or the interpersonal convergence of gestures (Kopp & Bergmann, in press). I will discuss empirical and modeling results indicating different mechanisms to be at work, e.g., automatic (low-level) and strategic (high-level). This suggests a need for dual-route models that come within reach based on the previously mentioned deep modeling attempts. The production model can account for coordination of gesture features in a top-down fashion, e.g. based on grounding; the perception model helps to explain how gesture features are affected by the mere observation of another’s gesture. The key question is how those automatic and strategic alignment effects interact in an integrated model.

Related to this is the question how speech-gesture behavior affects how we perceive and evaluate our human or artificial interaction partners. It is well established that non-verbal behaviors increase perceived naturalness and likability. Beyond this, I will present evaluation results (Salem et al. 2012) demonstrating how flexible and powerful models of multimodal speech-gesture behavior offer unique opportunities to fasten on the decisive effects of how multimodal ensembles are orchestrated.

2. ACKNOWLEDGMENTS

This research was partly supported by the Deutsche Forschungsgemeinschaft (DFG) within the Center of Excellence 277 „Cognitive Interaction Technology“ (CITEC) and the Collaborative Research Center 673 „Alignment in Communication“.

3. REFERENCES