Lifelike Gesture Synthesis and Timing for Conversational Agents

Ipke Wachsmuth
Artificial Intelligence Group
Faculty of Technology
University of Bielefeld
D-33594 Bielefeld, Germany
ipke@techfak.uni-bielefeld.de

Extended Abstract for GW2001 – Invited Lecture

Besides the inclusion of gesture recognition devices as an intuitive input modality, the synthesis of lifelike gesture is finding growing attention in human-computer interface research. In particular, the generation of synthetic gesture in connection with text-to-speech systems is one of the goals for embodied conversational agents which have become a new paradigm for the study of gesture and for human-computer interface [1]. Embodied conversational agents are computer-generated characters that demonstrate similar properties as humans in face-to-face conversation, including the ability to produce and respond to verbal and non-verbal communication. They may represent the computer in an interaction with a human or represent their human users as "avatars" in a computational environment.

In this context, this contribution focusses on an approach for synthesizing lifelike gestures for an articulated virtual agent, with particular emphasis on how to achieve temporal coordination with external information such as the signal generated by a text-to-speech system. The context of this research is the conception of an "articulated communicator" that conducts multimodal dialogue with a human partner in cooperating on a construction task.

Gesture production and performance in humans is a complex and multi-stage process. Abstract gesture specifications are generated from spatiotemporal representations of "shape" in the working memory on cognitively higher levels, and then transformed into patterns of control signals which can be interpreted by low-level motor systems; the resulting gesture exhibits characteristical shape and dynamic properties enabling humans to distinguish them from subsidiary movements and to recognize them as meaningful [2]. In particular, gestural movements can be considered as composed of distinct movement phases which form a hierarchical kinesic structure. In coverbal gestures, the stroke (the most meaningful and effortful part of the gesture) is tightly coupled to accompanying speech, yielding semantic, pragmatic, and even temporal synchrony between the two modalities [5].

Although promising approaches exist with respect to the production of synthetic gestures, most current systems produce movements which are only parametrizeable to a certain extent or even rely on predefined motion sequences. Thus achieving precise timing for accented behaviors in the gesture stroke as a basis to synchronize them with, e.g., stressed syllables in speech remains a research challenge. For instance, the GeSysCa system by Lebourque and Gibet [4] produces (French SL) sign language gestures from explicit representations, based on a limited set of motion primitives (straight, curved, and wave-like movements) that can be combined to more complex gestures and reproduce natural movement characteristics. However, adaptation of the movement's temporal and kinematic properties as required in coverbal gesture is out of their focus. The REA system by Cassell and coworkers (in [1]) implements an embodied agent which is to produce natural verbal and nonverbal outputs regarding various relations between the used modalities. In the gesture animation process, a behavior is scheduled that, once started, causes several motor primitives to be executed. The REA gesture model employs standard animation techniques, e.g. keyframe animation and inverse kinematics. Although the issue of exact timing of spoken and gestural utterances is targetted in their work, the authors state that it has not yet been satisfactorily solved.

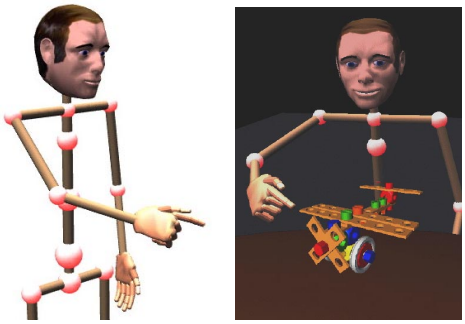

Fig. 1. Articulated Communicator

The general goal of our research on gesture synthesis is an operational model that enables real-time and convincingly lifelike gesture animation from representations of spatiotemporal gesture knowledge by way of a gestuary [2]. To this end, the model must provide sufficient means of motion representation, planning, and control in order to produce multiple kinds of gestural movements of a highly articulated figure as shown in Fig.1. Furthermore, it must be flexible in the sense that gestures can be parametrized with respect to kinematics, i.e. velocity profile and overall duration, as well as to shape properties.

We have developed a hierarchical model for planning and generating lifelike gestures which is based on findings in various fields relevant to the production of gesture in humans [3]. Our approach grounds on knowledge-based computer animation and encapsulates low-level motion generation and control, enabling more abstract control structures on higher levels. In summary, the gesture planner forms a movement plan, i.e. a tree representation of a temporally ordered set of movements constraints, by (1) retrieving a feature-based gesture specification from the gestuary, (2) adapting it to the individual gesture context, and (3) qualifying the movement constraints to the extent possible by temporal integration with external timing constraints. These techniques are used to drive the kinematic skeleton of a figure (see fig.1) which comprises 43 degrees of freedom (DOF) in 29 joints for the main body and 20 DOF for each hand. Gesture planning modules, including the gestuary, and the hand motor system were completed in an experimental implementation; the methods for trajectory formation are currently further elaborated. As the model is particularly conceived to enable natural cross-modal integration by taking into account temporal synchrony constraints, further work includes the integration of speech-synthesis techniques as well as run-time extraction of temporal constraints for the coordination of gesture and speech.

References

[1] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill (eds.). Embodied Conversational Agents. Cambridge (MA): The MIT Press, 2000.

[2] J.P. deRuiter. Gesture and Speech Production. PhD thesis, University of Nijmegen. MPI Series in Psycholinguistics, 1998.

[3] S. Kopp & I. Wachsmuth. A knowledge-based approach for lifelike gesture animation. In W. Horn (ed.) ECAI 2000 Proc.14th European Conference on Artificial Intelligence, Amsterdam: IOS Press, 2000.

[4] T. Lebourque & S. Gibet. A complete system for the specification and the generation of sign language gestures. In A. Braffort et al. (eds.), Proceedings International Gesture Workshop (pp. 227-238). Berlin: Springer-Verlag (LNAI 1739), 1999.

[5] D. McNeill. Hand and Mind: What Gestures Reveal about Thought. Chicago: University of Chicago Press, 1992.

Bio: Prof. Ipke Wachsmuth runs a lab at U of Bielefeld where multimodal interfaces are researched widely in the context of virtual reality. He is also co-director of the Collaborative Research Centre "Situated Artificial Communicators" (SFB 360) in Bielefeld which is supported by the Deutsche Forschungsgemeinschaft and which is the context of this work.