14.4.99 AA'99 Workshop Seattle

Natural Timing in Coverbal Gesture of an Articulated Figure

Stefan Kopp Ipke Wachsmuth

AG Wissensbasierte Systeme
Technische Fakultät, Universität Bielefeld
Postfach 100 131, D-33501 Bielefeld, Germany
e-mail: {skopp, ipke}@TechFak.Uni-Bielefeld.DE

1 Introduction

Anthropomorphic virtual agents have moved into the focus of human-computer interface researchers. Such agents mediate between the user and the technical system to establish a more intuitive communication link. Human communication is not restricted to spoken language as the only modality, but nearly always comprises additional coverbal utterances, i.e., gesture and mimics. It is natural to search for technical systems which can be instructed by the user using spoken language and gestures cooperatively. On the other hand, the technical system, too, should be able to produce proper multimodal output, in order to achieve greater comprehensibility in the opposite direction, as well.

In this paper, a virtual anthropomorphic agent is presented which is based on an articulated figure. Based on preceding work in this line [JW96], the aim of our current research is to make the agent able to perform coverbal gestures in a natural fashion. Besides the selection of the appropriate body movement to be performed as coverbal gesture, another crucial point is the correct timing of the motion with respect to the generated spoken utterance. The relationships between speech and gesture are discussed and consequences for generating coverbal gesturing and animation control of an articulated figure are drawn.

2 Speech, Gesture, and Timing

Evidence from many sources suggest a close relationship between speech and gesture. The first thing to notice is that people prefer to use speech and gesture simultaneously in face-to-face communication. Moreover, the sufficiency of a subject's utterance is considered to be maximal, if both, spoken language and gesturing are used in combination [HM93]. Kendon generally suggested that people tend to convey those concepts difficult to express in language by gesture [Ken86]. In this sense, the relationship between gesture and speech can be viewed as to resemble the interaction of words and graphics in the generation of multimodal output [FM91]. At the cognitive level, McNeill [McN92] proposed a concept of 'growth points' as the starting point of a microgenesis which leads to the integrated performance of speech and gestures. Different verbal utterances and accompanying gestures (i.e., iconic, metaphoric, beat, pointing) arise out of different growth points which could occur in a great variety.

Besides such theoretical and somewhat hypothetical approaches, many studies were carried out, which yield concrete results concerning the cooperative use of gesture and speech. McNeill pointed out that speech and gesture are expected to synchronize with respect to semantics and pragmatics. That is, co-occurring gesture and speech present the same meanings and perform the same pragmatic functions

[McN92]. Speech and gesture in automatically generated multimodal output hence have to be coordinated precisely in order to obtain an anthropomorphic agent capable of communicating understandably, believably, and naturally. To this end, the content of the actual verbal utterance, the discourse, as well as the possible meanings of the gesture have to be considered [JMN+94].

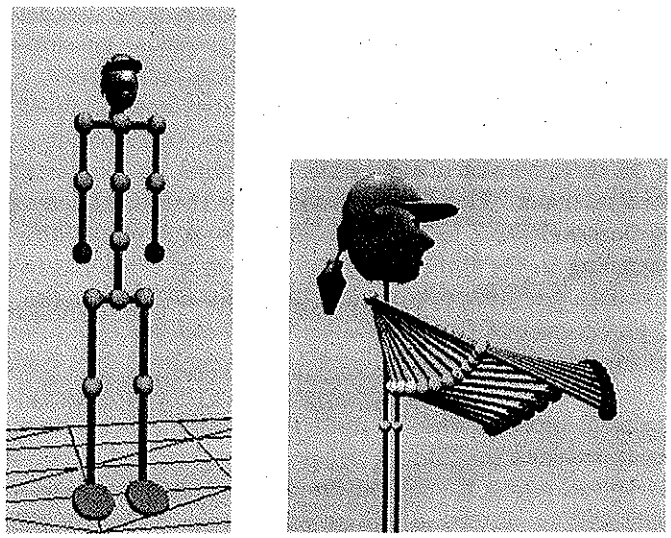
Besides the informational relationships between the two modalities, clearly, gesture and speech are related in time. Thus, coverbal gesturing of a virtual agent necessarily needs to be timed appropriately with respect to the generated speech, e.g. intonation, and vice versa. Consider for example the verbal utterance "put this wheel on that table". The appearance of a single accompanying pointing gesture provides additional spatial information. But whether this information concerns the wheel or the table depends solely on the timing of the gesture "stroke" (the moment of maximal expression of the gesture, normally the point of maximal jerk). McNeill investigated the question of when gestures are more likely to occur during spoken utterances in general [McN92]. As a result, gestures mostly occur with the specific sentence elements that convey new, and furthermore, most essential information with respect to the discourse.

With regard to the more exact temporal relationship between gesture and speech, it is established that a speaker's body is precisely synchronized with its own speech [Con86]. This synchronization holds across multiple levels and is assumed to be due to the same control mechanism underlying the global coordination of speech, as well as other coordinated rhythmic activity like hand movements [CP96]. When considering the progression of human gestures in time, it is observed that gestures are normally composed of "Gesture Phrases". These, in turn, consist of one or more movement phases, namely preparation, various holds, stroke, and retraction, forming a hierarchical kinesic structure of the gesture. The stroke, which is the obligatory part of a Gesture Phrase, is mostly preceded by a preparatory movement and followed by a retracting movement which either moves the limb back to a rest position or repositions it for the beginning of a new gesture phrase [Ken86]. According to Kendon, the gesture can be viewed as "nucleus" of movement having some definite form and enhanced dynamic qualities. Furthermore, he suggests that there can be found a close fit between the phrasal organization of gesticulation and the phrasal organization of speech. In particular, if the flow of speech is segmented into "Tone Units", there is usually a Gesture Phrase corresponding to each Tone Unit, both of them produced from the same underlying unit of meaning. Kendon points out that Gesture Phrases often begin in advance of the related Tone Unit and are often completed before the Tone Unit's completion.

On the lowest level of the kinesic hierarchy, it can be stated more precisely that the stroke never follows the nucleus of the Tone Unit, but is completed either before the nucleus, or just as its onset [Ken80]. Most preceding gestural strokes do so by at most one syllable's length. The preparation phase typically anticipates, by a brief interval, the linguistic segments that are coexpressive with the gesture's meaning. Butterworth and Beattie succeeded in quantifying the precendence for a certain kind of gestures [BB78]: The initiation of iconic gestures usually precedes and never follows the words with which they are associated, the mean delay being about 0.80 seconds, with a range of 0.10 to 0.25 seconds.

Little work has been done so far with respect to the automatic generation of convincing natural coverbal gesturing based on these theoretical findings. Cassell et al. [JMN+94] present a system that generates speech, intonation, and gesture in conversational interaction. Their virtual agent performs the gesture stroke on a certain ascociated word determined by the utterance's information structure, which defines its relation to other utterances in the discourse and to propositions in the relevant knowledge pool [Cas96]. Those gestures ascociated with a word are aligned with the stressed syllable of that word, i.e., beats. Other gestures which require

a preparatory phase start at the beginning of the intonational phrase in which the ascociated word occurs. Cassel et al. discovered that it is possible to specify quite successfully the temporal relationship between gesture and speech, i.e., when gestures might be expected in a discourse.
In summary, observations on the temporal relationship between gesture and speech support the view that gestures usually precede speech in a definite way. Moreover, different types of gestures may have distinct characteristics in form and motion, e.g. different dynamics, and hence have to be coordinated in different ways with the verbal utterance.



(a) (b)

Figure 1: *(a)* The skeleton of the articulated figure. *(b)* Multiple exposure of a pointing gesture. First the arm is moved upwards in preparation for the actual gesture stroke, which is downwards.

3 Articulated Pointing Gestures

Much of the research discussed in the previous section provides hints which could lead us to a predictive theory of coverbal gesture use. This in turn can be used for the gesture animation of a multimodal virtual agent. Gesture Animation therefore requires techniques for controlling the form and dynamics of its body movements, in order for the gesture to appear at the proper time in animation, as well as to provide natural expression. Thus, we consider the problem of generating gestures which satisfy given time constraints, such as starting point, end point, and moment of fullest expression, while maintaining natural appearance. Once body movements of the virtual agent can be adjusted and controlled exactly with respect to form and time, integration with speech flow and a natural timing of coverbal gesturing can be achieved more readily. In our current investigations, we are concentrating on coverbal pointing gestures which are, due to the necessity of exact timing, very challenging. Moreover, natural pointing gestures have, inspite of their great significance in human-computer interaction, found little attention so far. Pointing gestures, which have been considered in every existing classification, are well under-

stood as a classical triphasic (preparation, stroke, retraction) deictic movement to an object that is simultaneously referred to in the speech [RS91]. Whereas deictic gestures that appear during narratives rarely point to concrete entities, we consider here primarily pointing movements used by the anthropomorphic agent to select concrete objects in his virtual environment. That is, the agent is able to communicate definite spatial references by natural pointing movements which can be exactly timed to satisfy constraints imposed by simultaneous speech.

3.1 Animating the Articulated Figure

To generate natural gesturing we have started to model a virtual anthropomorphic agent and control the motions of its body. We therefore employ an articulated figure, inverse kinematics, and appropriate animation techniques.

Our virtual agent is structured as an articulated body defined by a skeleton [Alt98]. A skeleton is a connected set of segments, corresponding to limbs, and joints. Limbs, in turn, consist of rigid links and joints. A joint is a skeleton point where two links intersect. The links connected to that point may move, and the angle between the two links is called the joint angle. A joint can have at most three angles corresponding to the three rotational degrees of freedom. In order to allow only realistic human postures, appropriate limits for all joint angles are specified. In figure 1(a) an articulated figure is shown which consists of six limbs (right arm, left arm, right leg, left leg, spine, and neck-head complex), each composed of three links.

This articulated figure is able to perform a variety of pointing gestures, with respect to target positions in different distances, heights, and directions. In general, motion of the virtual agent is specified by defining motion of the skeleton, which in turn is generated by alteration of the joint angles. Thus, a given set of joint angle values prescribes a posture of the virtual body. Animations of the virtual body are defined by sequences of postures. For our figure, one just has to define significant postures for selected time points, which state the so called 'keyframes' of the animation. Then the system is able to find any joint angle at intermediate times by using interpolation methods. This technique, called 'parametric keyframing' [BW71], enables the figure to perform varying gestures such as pointing to arbitrary points in space, after adjusting the essential postures at gesturing time according to information like spatial deictic references. In figure 1(b) some postures of a pointing gesture are illustrated. Note the preparatory movement which first raises the arm. It is followed by the gesture stroke positioning the limb's end part in the correct direction and expressing the actual pointing.

To ease the definition of key postures, the discrete positions and orientations for the end parts of the limbs can be specified directly in the figure's workspace, i.e. cartesian space. Then inverse kinematics is used to compute automatically the necessary joint angles for other parts of the figure, in order to put the specified part in a desired position. This mapping from cartesian space into joint angle space is achieved by a recurrent network which uses several geometrical relations to approximate the same value several times in parallel [CS93]. Then, the mean value of these multiple computations is calculated (hence 'MMC net') and fed back for the next iteration. The net relaxes to adopt a stable state corresponding to a geometrically correct solution. By introducing constraints after the mean value has been computed, this approach can easily cope with limitations of the joint space, like those due to restrictions of the human body.

3.2 Natural Animation of Pointing Gesture

The findings in gesture research impose severe consequences for coverbal gesturing of an anthropomorphic agent. First, the movements of the agent's limbs have to depict gestures which must synchronize with speech semantically and pragmatically. To this end, the essential characteristics of a gesture, namely definite form and dynamic qualities, need to be recognizeably reproduced. In addition to the final shape of the motion of the articulated figure, we have to determine an appropriate timing for the movements, if they are to appear natural and should synchronize properly with spoken utterances. To this end, animation control has to take into account all movements phases of the lowest level of Kendon's kinetic hierarchy [Ken80]. For pointing gestures, the agent has to finish the gesture stroke with one arm pointing at the target position. This event must shortly precede or coincide with the stressed syllable of the associated word, which in most cases is the only temporal specification at disposal.

The pointing stroke, the moment of fullest expression of the gesture, originates from a short acceleration followed by a abrupt retardation, resulting in a significant jerk about the limb's end (see figure 2). In human pointing gestures this jerk sometimes can be observed solely with movement of the hand. But as our articulated figure currently lacks a proper model of the human hand, the stroke is performed mainly by the forearm (see figure 1 (b)), which conveys the gestural meaning as well.

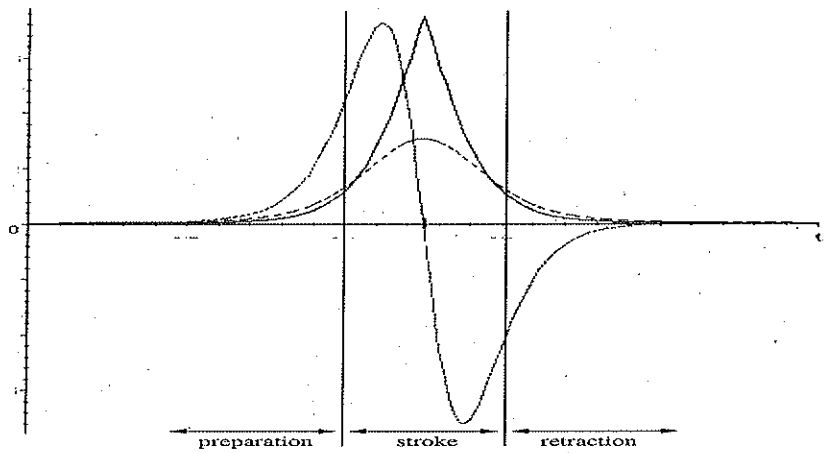


Figure 2: *(a)* Temporal characteristics of a complete pointing movement. The solid line shows a typical trajectory of the hand. In addition, the velocity profile (dashed) and acceleration (dotted) of the end-effector are drawn.

The preparation phase, which occurs on its own schedule, is crucial for the question of gesture timing, and can, as well as stroke-hold phases, compensate for mismatches of speech-gesture synchrony [McN92]. The agent adopts appropriate preparation and retraction movements in order to generate proper transitions between successive strokes. The resulting motion must satisfy all timing and posture constraints, due to required pointing strokes at given time points and desired target positions. Therefore, the relevant key postures of the predefined pointing sequences are adjusted accordingly. Simultaneously, first order discontinuities in the final motion of the limb have to be prevented to guarantee natural and fluid appearance. Thus, the method of interpolation between the chosen key postures is taken into

account, because motion kinematics can be modified by way of calculating a different number of postures in between, as well as altering the parameter variations between two successive postures.

After pointing, humans tend to retract the limb back to a comfortable rest position. However, if the next movement is expected to appear immediately or within a short time interval, both strokes are incorporated into a motion along the shortest path between the start positions for the stroke. Experience in computer animation has shown that motion which minimizes energy looks natural [RGBC96]. Therefore, generation of realistic motion transitions ultimately have to rely on minimization of cost functions leading to appropriate preparation and retraction phases.

Finally, natural motions do not simply start and stop suddenly. Rather, they accelerate from a starting point and decelerate to a stop. This is due to moment of inertia of the body segments and the forces which cause the movement. In classical computer animation, this effect is known as "slow-in and slow-out" and is usually added by the animator manually [Mae96]. In our articulated figure, which is assumed to be visualized with a fixed frame rate, the kinematics (velocity and acceleration) of the motion is generated by adjusting the amount each joint angle is incremented or decremented for the next posture. Greater variations cause the motion to appear faster; smaller alterations seem to lower the speed. Thus, in order to achieve slow-in and slow-out, each variation is calculated depending on the position of the current posture within the course of a total gesture sequence. The resulting velocity profile of a complete pointing motion, including acceleration from rest position and final decelaration, is illustrated in figure 2.

4 Conclusion

In this paper, an articulated agent was presented which was conceived to enable coverbal gesturing. The relationships between speech and gesture were discussed and consequences for generating coverbal gesturing and animation control of the articulated figure were drawn. In the current state of our demonstrator, the figure can perform pointing gestures to arbitrary target positions. The individual gestural phrase, including a preparation, stroke, and retraction phase, can be fitted in a fixed time frame. Most important, our system performs temporal planning with respect to the number of intermediate animation frames in order to meet time constraints for the overall gestural phrase. Hence, the agent is able to express the major stroke at a given time point. Ongoing work is directed to incorporate successive pointing gestures in homogeneous and natural body movements.

The anthropomorphic agent, at present, is employed as an interface agent in a scenario of virtual construction where it serves to communicate spatial references by pointing to locations. Our mid-range goal is the automatic generation of natural coverbal gestures by the articulated figure, so it can be enabled to carry simple multimodal dialogs, e.g., in answering questions about the locus of objects, including deictic references.

References

[Alt98] Frank Althoff. Entwicklung eines Basissystems für die Modellierung und interaktive Simulation einer artikulierten Figur. Master's thesis, Faculty of Technology, University of Bielefeld, Bielefeld, Germany, 1998.

[BB78] B. Butterworth and G. Beattie. Gesture and silence as indicators of planning in speech. In P.T. Smith R.N. Campbell, editor, *Recent advances in the psychology of language*, pages 347–360. Plenum, New York, 1978.

[BW71] N. Burtnyk and M. Wein. Computer-generated key-frame animation. *Journal SMPTE*, 80:149–153, 1971.

[Cas96] Justine Cassell. Believable communicating agents. SIGGRAPH '96 Course Note, 1996.

[Con86] William S. Condon. Communication: Rhythm and structure. In J. Evans and M. Clynes, editors, *Rhythm in Psychological, Linguistic, and Musical Processes.* Thomas, Springfield, IN, 1986.

[CP96] Fred Cummins and Robert Port. Rhythmic commonalities between hand gestures and speech. In *Proceedings of the Cognitive Science Society Annual Meeting 1996*, University of California, San Diego, 1996.

[CS93] H. Cruse and U. Steinkühler. Solution of the direct and inverse kinematic problems by a common algorithm based on the mean of multiple computation. *Biol. Cybern.*, pages 345–351, 1993.

[FM91] S. Feiner and K. McKeown. Automating the generation of coordinated multimedia explanations. *IEEE Computer*, 24(10), 1991.

[HM93] Alexander G. Hauptmann and Paul McAvinney. Gestures with speech for graphic manipulation. *International Journal of Man-Machine Studies*, 38:231–249, 1993.

[JMN+94] J. Cassell, M. Steedman, N. Badler, C. Pelachaud, M. Stone, B. Douville, S. Prevost, and B. Achorn. Modeling the interaction between speech and gesture. In *Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society*, 1994.

[JW96] Tanja Jörding and Ipke Wachsmuth. An anthropomorphic agent for the use of spatial language. In *Proceedings of ECAI'96-Workshop "Representation and Processing of Spatial Expressions"*, pages 41–53, Budapest, 1996.

[Ken80] Adam Kendon. Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key, editor, *The Relationship of Verbal and Nonverbal Communication*, pages 207–227. The Hague, Mouton, 1980.

[Ken86] A. Kendon. Current issues in the study of gestures. In Nespoulous, Perron, and Lecours, editors, *The Biological Foundations of Gestures: Motor and Semiotic Aspects*, pages 23–47. Lawrence Erlbaum Associates, Hillsday N.J., 1986.

[Mae96] George Maestri. *Digital Character Animation.* New Riders Publ., Indianapolis, IN, 1996.

[McN92] Davic McNeill. *Hand and Mind: What Gestures Reveal about Thought.* University of Chicago Press, Chicago, 1992.

[RGBC96] Charles Rose, Brian Guenter, Bobby Bodenheimer, and Michael F. Cohen. Efficient generation of motion transitions using spacetime constraints. In *Proceedings of the 23rd annual conference of Computer graphics*, pages 147–154. ACM SIGGRAPH, ACM Press, 1996.

[RS91] Bernard Rime and Loris Schiaratura. Gesture and speech. In R. S. Feldman and R. Rime, editors, *Fundamentals of Nonverbal Behavior.* Press Syndicate of the University of Cambridge, New York, 1991.