Let’s shake hands!
On the coordination of gestures of humanoids

Zsófia Ruttkay and Herwin van Welbergen

1. INTRODUCTION

Hand gestures are important means of expressivity for humanoids. In this paper by humanoid we understand user-controlled or autonomous virtual humans (VHs) or conversational agents (ECAs) [4] as well as human-like robots [18]. We cover both cases of human-humanoid and humanoid-humanoid interaction. The semantics, the morphology, the variations in performance of gestures reflecting cultural, affective and other characteristics of the speaker [8]. as well as general gesture movement laws [6] have been addressed Our focus in this paper is the issue of *coordination* of hand gestures to external signals. One type of coordination, alignment of speech-accompanying gestures to the speech, has been studied extensively, and different design principles have been formulated and implemented for specific applications with virtual humans [11, 13, 25]. In these cases, the phonological synchrony rule [15] have been taken as basis, usually resulting in gestures timed to the speech – even if it is generated by TTS. An exception is [24], where in assembly tasks, where a physical manipulation may be accomplished in a shorter or longer time, the speech is aligned to the manipulative hand gestures. Another domain where two-handed gestures play a role is sign language [9]. Also, mechanisms for fast planning for deictic gestures have been proposed [14]. Our ongoing research extends these works in the following aspects:

- We propose a coordination scheme which is *more general*, allowing to take into consideration external events such as tempo indication or perceived state information about the interlocutor of the ECA.
- We allow the declaration of coordination requirements on a *low level of granularity*, looking at different stages of gestures. Such a refined approach makes it possible to perform experiments on e.g. expressivity and style, and to include timing strategies as a means to fine-tune the gesturing behavior of a humanoid.
- Our main interest is in *reactive scheduling and planning of gestures* with reference to an environment influencing the timing of the gestures.
- We are using the (still under development) *BML language* for the formulation of scheduling requirements. As BML is meant to become a general-purpose markup language [3], our testing and extension of its constructs contributes to the development of this unifying language.

One may wonder if it is necessary to endow humanoids with the capability of such subtle coordination. What are the application contexts where such coordination is needed?

In present applications typically an ECA is either ‘alone’, or at least not paying attention to the (real or virtual) partner while talking. If he does, it is via eye contact, and not body contact. Indeed, modeling gaze behavior [1] during conversation addresses a similar, albeit much simpler coordination problem, where no physical constraints are present to influence the gazing behavior of the parties.

Current humanoids are not adaptive and robust enough in their reactive gesturing behavior. However, consider a virtual world inhabited by multiple humanoids, either autonomous or as avatars driven by real people’s intentions. A very natural ‘act’ in such environments too is greeting, to initiate conversations, or indicate (e.g. in games, simulations) their relationship. Quite long ago hand shaking was one of the ‘wishes’ formulated as something a humanoid should be capable of [22]. Yet, it has not become a practice. Also, the subtle, reactive coordination assumes (real or simulated) perception of humanoids – a topic which is getting more attention recently [20].

Though our own field is virtual humans, we emphasize that a subtle, reactive coordination scheme can be applied for robots, where physical contact is a plus dimension of communication [16, 19], especially for building a common ground and expressing emotions [18, 3].

HMI, University of Twente, e-mail: zsofi@cs.utwente.nl

In our paper first we give an ontology of types of coordination problems related to hand gesturing. Then we demonstrate the problems on the basis of two examples: clapping as a rhythmic 2-handed gesture; and hand-shaking which is a single-handed gesture of two persons, which 'makes sense' only if the required coordination takes place. Hand gestures like these fall outside the usual speech-driven hand gesturing of humanoids, both concerning the coordination problems involved and the application context. Then we explain our own multimodal planning system to deal with the subtle coordination problems. Finally we outline the extra requirements these coordination tasks pose on the BML language. Our work is done parallel to gathering empirical data on relevant human-human gesturing, by recording and analyzing captured motion data, and possibly also other nonverbal signals gained from the recorded video. The proposed coordination strategies are to be demonstrated with virtual humans, and their gestures are to be tested. The present work is a step towards our ultimate goal; that is to build a multi-layer behavior engine for virtual humans, particularly for serious game applications.

2. ONTOLOGY OF COORDINATION OF HAND GESTURES

In order to characterize the coordination problems related to hand gestures, the following aspects need to be specified:

1. Origin of signals involved in the coordination
   a. an event or signal of the world;
   b. one or more other modalities of the humanoid self (speech, other hand, gaze,…);
   c. one or more modalities of another humanoid or a real human.
2. The flexibility of the timing of signals involved in the coordination
   a. a signal is an inflexible signal, if its given timing cannot be changed;
   b. the timing of the signal is flexible within certain constraints.

Table 1 gives an overview of the 3x2 cases, one of which – flexible signals from the word – is empty.

The concept of an inflexible signal is crucial: that is the signal which is to be taken 'as is', no alteration of its given timing is possible. In general, 'signals of the world' are such: one cannot change the tempo of a recorded music or the trajectory of a ball to be caught (but you may apply different strategies to catch it). In speaker-listener situation, the listener's feedback signals are the flexible ones, to be aligned with the speaker. (It is another question that the speaker may interpret the feedback so that he needs to talk slower, as the listener cannot follow his fast speech.)

But in cooperative tasks, both parties are flexible in order to achieve the common goal. On the other hand, from time to time one of them takes the initiative and expects the other to comply. Hence it may change from time to time if one or the other's hands (or other modality) is the leading, non-flexible signal.

Flexibility **Origin**	**Inflexible**	**Flexible**
World	- pointing at a *moving object* - clapping to *rhythm of music*	-
Humanoid's own modality	- gesture aligned to *speech* which is taken as leading signal	- *gaze and hand* coordination
Other humanoid's modality	- back-channeling as listener to a *speaker* e.g. by head nods	- *hand shake* - *two hands* involved in taking over an object

Table 1: Overview of signals for gesture coordination, the signal categorized is shown in italics.

Our earlier work on a listening conductor [2] made it clear that even a conductor-conducted musician relationship, it is not always the conductor who's hand movements are the leading signals – occasionally, he must adjust his conducting to the (too slow) music produced by the player. A similar, subtle 'game' happens often, albeit usually in an unconscious way, in case of interpersonal manipulative or communicative hand gestures, like carrying a heavy bag together with somebody, or shaking hands to greet.

3. COORDINATION IN TWO EXAMPLES

In order to perform a gymnastic exercise in a given rhythm, a hand gesture like a clap above the head is to be repeated according to the (may be changing) tempo 'dictated' by music, or a metronome. In this case, the metronome or music is a fixed signal of the world, and the hand gesture of the humanoid is flexible, which needs to be synchronized to tempo. The freedom in the synchronization is in how the total time is to be distributed among stages of the hand gesture. Note that in physical exercises it is important to be specific about the scheduling of the phases of the gesture, demanding e.g. that the stroke part of the gesture is done much faster, the continuity of the motion requires that no hold times are used, etc. We have gathered mocap data of joints and video and audio, and analyzed the synchronization strategies for tempo changes [23]. In a nutshell, we found that:

- The phonological synchrony rule was valid for counting while clapping.
- The clapping movement is often sped up just by decreasing the path distance decreases linearly with the clapping speed.
- A pre-stroke hold can be used as a slowdown strategy.
- The standard deviation of the relative phase angle between the left and right hand increased with the clapping frequency. No significant increase of the mean was observed.
- For our right handed subjects the motion was asymmetrical, the right hand was moving ahead in phase compared to the left.
-

Another example is the hand shaking of two (virtual) humans. There is little literature about how hand shake takes place among humans. Some studies address the variety in greeting gestures, depending on the social, gender and cultural characteristics of the people meeting [5, 7] and coordination in social interaction, in general [10]. As of the 'Western handshake', some normative guidelines are available for every-day scenarios [26]. The importance of establishing gaze contact, the strength of the grasp applied, the duration and number (2-4) and performance characteristic of 'pumping motions' such as to be performed from the elbow, are mentioned. Some social connotations of coordination and timing are noted. Particularly, the person initiating the hand shake is seen as the more dominant, socially higher-ranked - in Western business-like situations.

The coordination of this common greeting act is, in fact, a subtle process, where both participants are involved. One of the parties takes the initiative, and extends his right hand, to the 'normal' position of hand shaking start. If the other is to accept the hand, then he reaches towards the hand of the other. When does this movement start? What should be the target of the hand of the interlocutor? If the partner's hand is already in the hand shake start position in front of him, then obviously, the partner's hand is the target. But if the interlocutor started to move his hand while the partner's hand is still in motion (which is often the case in real life), the movement of the partner's hand is observed unconsciously: you do not grab a hand still in acceleration, but adapt your hand's motion in way that you both 'arrive' at a point where the hands hardly move and the palms are close enough to be embraced by the other's fingers. Once the hands are joined, they may kept sill, or a few 'pumping' motions take place, and at some point the parties extend their fingers and thus each of them may withdraw his hand. In reality, important factors of the entire process (who should start the hand shake, how long should it take) are controlled by social protocols and by visual and haptic feedback. The latter are used to accommodate to the special geometrical or other characteristics of the partner, such as size or position (e.g. seated).

The hand shake may be coordinated with other modalities: as soon as the hand contact is established, gaze contact should be established too, and kept as long as the hands are joined. Sometimes the partner does not respond to a hand shaking initiative, or keeps the hand beyond the will of the other. In the first case, how long should a humanoid wait with his extended hand to be grabbed by the partner? In the latter case, how to escape from such a situation? In reality, application of force may be enough, however, usually some social protocol is applied. The too long kept hand may be interpreted as sign of extra interest, or establishment of power relationship, and may be acknowledges or refused by a new communicative action (speech), also with the goal of ending the hand shaking.

Currently, we are busy with recording mocap and video data in situations where two persons need to great each other by handshake in a spontaneous, natural way; and in situations where one of the parties (an experimenter) tries to influence the handshake in different ways, e.g. being too slow with response (including no response at all), influencing the number of pumping motions and the duration of holding the other's hand. Besides eliciting motion characteristics, we are looking at the timing of gaze behavior.

4. REPRESENTATION AND PLANNING OF SYNCHRONIZATION

BML is a multimodal generation language, describing synchronization between speech and animation on such a level that it can be used as input for the final process of multimodal generation [11]. We extended BML, e.g. with *Observers* to monitor *outside actions*, that is ones not related to the humanoid's own modalities and synchronize to those too.

We have developed an interactive environment where the user may specify explicitly the tempo of repetitive gestures like claps, and tell the tempo of preparation and stroke, and how to distribute the remaining time (if any) between holds before and after the clap stroke. We implemented a demonstrator where a virtual character performs the clap sequence according to the specification, see figure 1.

Besides direct timing prescriptions, the amplitude of the motion may be prescribed too, which has consequences on timing. The amplitude-duration of stroke relationship is based on empirical tests with humans. Such a constraint between amplitude and duration is considered as characteristic of the clap as a gesture. However, when planning clapping, this constraint is to be taken into account in addition to the explicitly declared constraints.

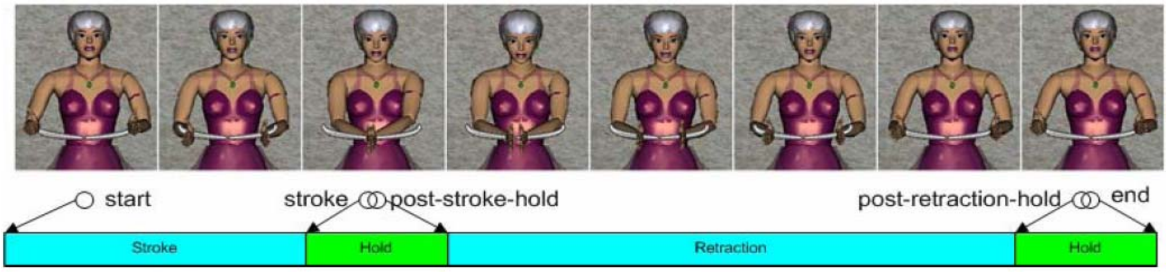

Figure 1: Duration of stages of a clap.

Currently, there is no built-in mechanism to prevent or correct inconsistent, infeasible specifications. Partial (underconstrained) specifications will be planned according to a 'default' strategy. E.g. if the tempo is to be slowed down, the duration of preparation and stroke will be slowed down proportionally, as default. However, the user may specify, for instance, that in the slow tempo the stroke should not be too slow, and the remaining time is assigned automatically to a hold.

For the clap gesture we used alignment points, identifying phases of hand gestures as the start and end of the entire gesture, the stroke, the pre- and post-stroke-hold, and retraction.

The timing constraints are expressed in BML linking (some of) these alignment points to each other and/or to the rhythmic signal of the world produced by a metronome. As a result of planning, all the alignment points of the subsequent claps get mapped to real time values, and the resulting animation is produced by the low-level animation engine which, in our current demonstrator, time-warps a default clap animation of the given amplitude.

5. DISCUSSION

The close look at clapping as (metronome) signal-driven inter-personal two-handed, and hand-shaking as a inter-personal cooperative gesture serve as studies for possible other cases where the success of some action is based on coordinated hand movement. We left unspecified in our framework how the signal from the outside world or the 'other' party is perceived and interpreted. In case of handshake, peripheral computer vision, or, if applied for robots, tactile feedback would be needed. Within our planning framework, it remains an issue how to propagate back updated, perceived information from the low-level planner to the higher level gesture scheduler?

Another interesting and difficult issue is the cognitive aspects of multi-party hand gesture performance. As mentioned in the introduction, in real life it depends on social and other characteristics of the partners who the 'leading' party will be. Also, deviations from a 'standard' performance convey meaning, intentions and believes of each party with respect to the other. Hence the subtle, perceived characteristics of the performance of a cooperative gesture may have effect on the cognitive aspects, as emotional state, believes, goals, user modeling. This issue has been raised at the recent BML meeting in a specific form, namely how to prepare a humanoid for failure in gesturing, e.g. due to malfunctioning of the lowest-level animation mechanisms. The issue of gapping the low-level gesturing and the mind via a feedback mechanisms is essential for interpreting behavior of partners of humanoids, and learning about the (potential) communicative partners.

REFERENCES

1. Argyle, M. and M. Cook: Gaze and mutual gaze, Cambridge University Press, 1976.
2. P. Bos, D. Reidsma, Zs. Ruttkay and A. Nijholt: Interacting with a virtual conductor. Proc. of 5th International Conference on Entertainment Computing, Cambridge, UK (September 2006), no. 4161 in Lecture Notes in Computer Science, Springer Verlag. 2006, pp. 25–30.
3. C. Breazeal: Social interactions in HRI: the robot view, . Systems, Man and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 34:2, 2004, pp. 181-186.
4. Justine Cassell: Embodied Conversational Agents. The MIT Press, April 2000.
5. A. Duranti: Universal and Culture-Specific Properties of Greetings, Journal of Linguistic Anthropology, Vol. 7, No. 1, 1997, pp. 63-97.
6. S. Gibet, J.-F. Kamp and F. Poirier: Gesture Analysis: Invariant Laws in Movement. In Gesture-based Communication in Human-Computer Interaction, LNCS/LNAI, Volume 2915, 2004, pp 1-9.
7. Paul E. Greenbaum and Howard M. Rosenfeld: Varieties of touching in greetings: Sequential structure and sex-related differences, Journal of

Nonverbal Behavior, Volume 5, Number 1, 1980. pp. 13-25.

8. B. Hartmann, M. Mancini, C. Pelachaud, Implementing Expressive Gesture Synthesis for Embodied Conversational Agents, Gesture Workshop 2005, LNAI, Springer, 2005.

9. Matt Huenerfauth: Representing Coordination and Non-Coordination in American Sign Language Animations. Behaviour & Information Technology, Volume 25, Issue 4, 2006, pp. 285-295.

10. A. Kendon: Movement Coordination in Social Interaction: Some examples described. Acta Psychologia, 32, 1970. pp. 100-125.

11. B. Krenn, H. Pirker: Defining the Gesticon: Language and Gesture Coordination for Interacting Embodied Agents, Proc. of the AISB-2004 Symposium on Language, Speech and Gesture for Expressive Characters, Convention of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour, University of Leeds, UK, 2004, pp.107-115.

12. S. Kopp, B. Krenn, S. Marsella, A. N. Marshall, C. Pelachaud, H. Pirker, K. R. Thórisson, H. Vilhjálmsson: Towards a Common Framework for Multimodal Generation: The Behavior Markup Language. Proc. of IVA 2006, pp. 205-217.

13. Stefan Kopp and Ipke Wachsmuth: Model-based animation of coverbal gesture. In Proceedings of Computer Animation, Washington, IEEE Computer Society, 2002, pp. 252-257.

14. J. Lester, J.Voerman,, S.Towns, C.Callaway: Deictic believability: Coordinated gesture, locomotion and speech in lifelike pedagogical agents, Applied AI, Vol. 13. No. 4/5. 1999. pp. 383-414.

15. D. McNeill: Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press, Chicago, 1995.

16. C. L. Nehaniv, K. Dautenhahn, J. Kubacki, M. Haegele, C. Parlitz, and R. Alami: A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction. Proc. of ROMAN 2005, pp. 371- 377.

17. Z. M. Ruttkay, J. Zwiers, H. van Welbergen, and D. Reidsma. Towards a reactive virtual trainer. In Proceedings of the 6th International Conference on Intelligent Virtual Agents, Springer, 2006. pp. 292–303.

18. Sidner, C. and C. Lee and C. Kidd and N. Lesh: Explorations in Engagement for Humans and Robots, Artificial Intelligence, May 2005.

19. P. Olivier, D.G. Jackson & C. Wiggins.A Real-world Architecture for the Synthesis of Spontaneous Gesture, Proceedings of 19th annual conference on Computer Animation and Social Agents (CASA2006), Geneva, Switzerland, 2006.

20. C. Peters: Evaluating perception of interaction initiation in virtual environments using humanoid agents. Proceedings of the 17th European Conference on Artificial Intelligence, 2006. pp. 46--50.

21. Michihiko Shoji, Kanako Miura, and Atsushi Konno: U-Tsu-Shi-O-Mi: The Virtual Humanoid You Can Reach, SIGGRAPH 2006 Emerging technologies

22. D. Thalmann, R. Boulic, Z. Huang, and H. Noser, Virtual and Real Humans Interacting in the Virtual World, Proc. International Conference on Virtual Systems and Multimedia `95, pp.48-57.

23. H. Van Welbergen and Zs. Ruttkay: On the parameterization of clapping, Proc. of Gesture Workshop 2007, Lisbon, Portugal, to appear.

24. I. Voss and I. Wachsmuth: Anticipation in a VR-based anthropomorphic construction assistant. In J. Jacko and C. Stephanidis (eds.): Human-Computer Interaction, Theory and Practice *(Part I)*, London: Lawrence Erlbaum Associates. 2003, pp. 1283-1287.

25. I. Wachsmuth and S. Kopp: Lifelike Gesture Synthesis and Timing for Conversational Agents. In I. Wachsmuth and T. Sowa (eds.): Gesture and Sign Language in Human-Computer Interaction Berlin: Springer (LNAI 2298), 2002, pp. 120-133.

26. Marta Wilson and Sharon Flinder: The Business Protocol Advantage, http://transformationsystems.com/Assets/BusProtocol.pdf