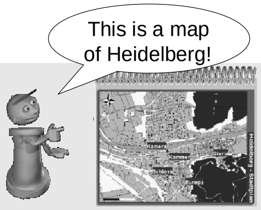
Gesture-Speech Interaction in the SmartKom Project

Antje Schweitzer and Grzegorz Dogil
Institute of Natural Language Processing
University of Stuttgart, Germany
{schweitzer,dogil}@ims.uni-stuttgart.de

Peter Poller,
DFKI, German Research Center for Artificial Intelligence
Saarbrücken, Germany
poller@dfki.de



The SmartKom project

The goal of the project is to create an intuitive multimodal dialog system that combines speech, gesture and mimics input and output. Interaction with the system follows the „situated delegation-oriented dialog paradigm“: user tasks are delegated to a virtual communication assistant, who elaborates and carries out the task, possibly in interaction with the user (Wahlster et al. 2001). The communication assistant is realized as a life-like character (named “Smartakus”) and presents the system output both visually and acoustically.

The project is funded by the German Ministry of Education and Research (BMBF) for the period of September 1999 - August 2003.

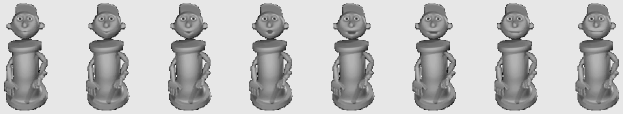
For further information see <u>*http://www.smartkom.org*</u>

Particular requirements of multimodal output in SmartKom

- Information conveyed by speech may be presented in another modality at the same time. This will influence **prosody**.
- Smartakus requires **lip synchronization**.
- When Smartakus gestures while speaking, **gesture-speech alignment** is necessary.

Lip synchronisation

To achieve high naturalness, perfect synchronization is crucial. Lip movement is generated by synchronized visual concatenation of appropriate mouth position pictures (representing so-called visemes) for each phoneme, yielding smooth lip movements (average 12 frames/sec). Visemes vary in jaw opening and lip rounding. We distinguish 8 visemes: rounded or unrounded, with 4 opening degrees.



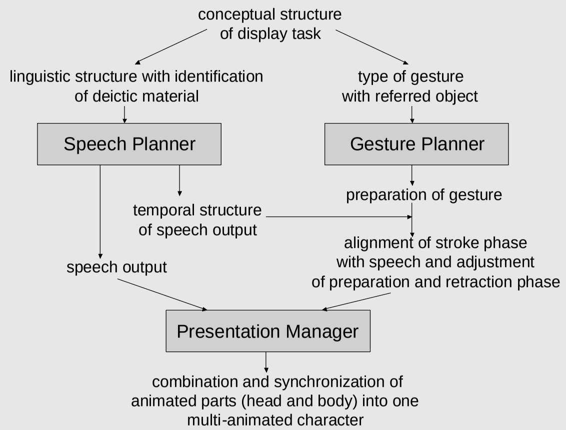
Coarticulation effects: Vowels are specified for both lip rounding and opening degree. Consonants are underspecified with respect to lip rounding, and consonants with posterior articulation places are only specified for a maximal opening degree. Thus, partially underspecified lip constellations for consonants can be realized by different visemes depending on the context.

Gesture-speech alignment

We distinguish between idle gestures, which occur in phases with no speech interaction, and meaningful gestures which accompany speech. The latter are mostly deictic gestures occurring during the presentation of graphical objects on the display; lexicalized gestures like nodding, shrugging etc. are also possible.

Speech-related gestures have to be aligned with the corresponding speech. According to the literature (McNeill 2000), gestures can be divided into three phases: preparation phase, stroke phase, and retraction phase. The stroke phase is what we perceive as the „meaningful“ core part of the gesture and is usually temporally aligned with the corresponding linguistic material, although the stroke phase may have a longer or shorter duration than the relevant speech material. If the stroke phase is shorter, preparation or retraction phase can be prolonged accordingly.

Gestures during speech add a new aspect to lip synchronization. Rendering of visemes and gestures is done off-line because of the complexity of the underlying 3D model. With gestures and lip movements occurring simultaneously, the number of all possible combinations of lip movements, head angles and body movements is too large to be prepared beforehand. Therefore, head and body are animated separately.



Effects of gestures on prosody

In SmartKom, all gestures accompanied by speech are deictic gestures. A pilot study shows that the corresponding deictic speech items are prosodically prominent.

Setup: A short dialog containing deictic pronouns was read by 27 professional and semi-professional speakers. We evaluated three sentences containing a deictic pronoun (i) in phrase-initial position substituting an accusative object („den“, *this one*), (ii) in phrase-initial position as a prepositional phrase („da“, *there*), and (iii) in phrase-final position as a prepositional phrase („hier“, *here*). The dialog contained explicit indication of pointing gestures accompanying these items.

Results: Deictic pronouns are prosodically more prominent than other pronouns. The phrase-initial ones were at least marked by a falling or rising pitch accent in almost all cases (37 out of 47), and often by strong rise-falls or by prosodic boundaries resulting in intermediate phrases containing only the deictic pronoun (9 out of 22 for (i), and 12/25 for (ii)). The results for the phrase-final pronoun „hier“ (iii**)** were less clear; prosodic boundaries were used in only 4 cases (4/25), and there were no rise-falls. However, falling accents were observed in most cases (21/25), often even combined with a pitch accent on the directly preceding word.

Conclusion

When aligning gestures and speech, prosody generation takes into account deixis. Beyond that, no explicit temporal modification of speech is necessary to adjust speech output to the duration of the gesture; instead, preparation or retraction phase can be modified according to the temporal structure of the accompanying speech.

For lip synchronisation, the concatenation of animated sequences rendered off-line yields smooth lip movements. Modeling of coarticulation effects enhances the naturalness of the animation.

References

- David McNeill (ed.), Language and gesture, Cambridge University Press, 2000.
- W. Wahlster, N. Reithinger, A. Blocher, SmartKom: Multimodal Communication with a Life-Like Character, Proceedings of Eurospeech 2001, Aalborg, Denmark