Multimodal interaction control in the MonAMI Reminder

Gabriel Skantze
Dept. of Speech Music and Hearing
KTH, Stockholm, Sweden
gabriel@speech.kth.se

Joakim Gustafson
Dept. of Speech Music and Hearing
KTH, Stockholm, Sweden
jocke@speech.kth.se

Abstract

In this demo, we show how attention and interaction in multimodal dialogue systems can be managed using head tracking and an animated talking head. This allows the user to switch attention between the system and other humans. A preliminary evaluation in a tutoring setting shows that the user's attention can be effectively monitored with this approach.

1 Introduction

Most spoken dialogue systems are based on the assumption that there is a clear beginning and ending of the dialogue, during which the user pays attention to the system constantly. However, as the use of dialogue systems is extended to settings where several humans are involved, or where the user needs to attend to other things during the dialogue, this assumption is obviously too simplistic (Horvitz et al., 2003). When it comes to interaction, a strict turn-taking protocol is often assumed, where user and system wait for their turn and deliver their contributions in whole utterance-sized chunks. If system utterances are interrupted, they are treated as either fully delivered or basically unsaid.

In this demo, we show how attention and interaction in multimodal dialogue systems can be managed using head tracking and an animated talking head. This allows the user to switch attention between the system and other humans, and for the system to pause and resume speaking.

2 The MonAMI Reminder

This study is part of the 6th framework IP project MonAMI. The goal of the MonAMI project is to develop and evaluate services for elderly and disabled people. Based on interviews with potential users in the target group, we have developed the MonAMI Reminder, a multimodal spoken dialogue system which can assist elderly and disabled people in organising and initiating their daily activities (Beskow et al., submitted). Information in their personal calendars can be added using digital pen and paper, allowing the user to continue using a paper calendar, while the written events are automatically transferred to a backbone (Google Calendar). The dialogue system is then used to get reminders, as well as to query and discuss the content of the calendar.

The system architecture is shown in Figure 1. A microphone and a camera are used for system input (speech recognition and head tracking), and a speaker and a display are used for system output (an animated talking head). As can be seen in the figure, all system input and output is monitored and controlled by an Attention and Interaction Controller (AIC). The purpose of the AIC is to act as a low level monitor and controller of the system's speaking and attentional behaviour. The AIC uses a state-based model to track the attentional and interactional state of the user and the system. The system is initially in a non-attentive state, in which the animated head looks down. As the user starts to look at the system, the animated talking head looks up and the system may react to what the user is saying. If the user looks away while the system is speaking, the system will pause and resume when the user looks back. If the user starts to speak while the system is speaking, the controller will make sure that the system pauses. The system may then decide to answer the new request, simply ignore it and resume speaking (e.g., if the confidence is too low), or abort speaking (e.g., if the user told the system to shut up).

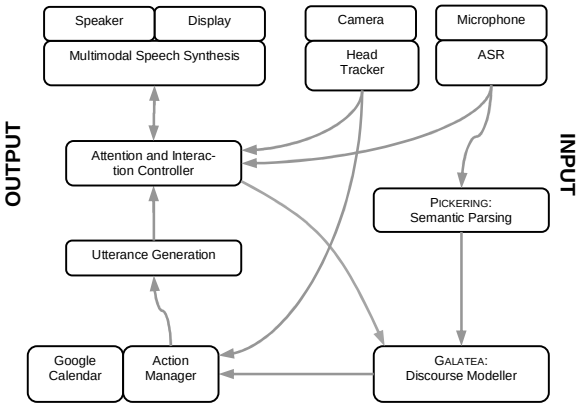

Figure 1. The system architecture in the MonAMI Reminder.

3 Preliminary evaluation

In the evaluation, we not only wanted to check whether the AIC model worked, but also to understand whether user attention could be effectively modelled using head tracking. Similarly to Oh et al. (2002), we wanted to compare "look-to-talk" with "push-to-talk". To do this, we used a human-human-computer dialogue setting, where a tutor was explaining the system to a subject (shown in Figure 2). Thus, the subject needed to frequently switch between speaking to the tutor and the system. A second version of the system was also implemented where the head tracker was not used, but where the subject instead pushed a button to switch between the attentional states (a sort-of push-to-talk). 8 subjects were used in the evaluation, 4 lab members and 4 elderly persons in the target group (recruited by the Swedish Handicap Institute).

An analysis of the recorded conversations showed that the head tracking version was clearly more successful in terms of number of misdirected utterances. The subjects almost always looked at the addressee in the head tracking condition, and did not start to speak before the animated head looked up. When using the push-to-talk version, however, they often forgot to "turn it off", which resulted in the system interpreting utterances directed to the tutor and started to speak when it shouldn't. The addressee of the utterances in the push-to-talk condition was correctly classified in 86.9% of the cases, as compared with 97.6% in the look-to-talk condition.

Figure 2. The human-human-computer dialogue setting used in the evaluation. The tutor is sitting on the left side and the subject on the right side

These finding partly contradict findings from previous studies, where head pose has not been that successful as a sole indicator for the addressee (cf. Bakx et al., 2003; Katzenmaier et al., 2004). One explanation for this might be that the subjects were explicitly instructed about how the system worked. Another explanation is the clear feedback (and entrainment) that the agent's head pose provided.

Acknowledgements

This research is supported by MonAMI, an Integrated Project under the European Commission's 6th Framework Program (IP-035147), and the Swedish research council project GENDIAL (VR #2007-6431).

References

Bakx, I., van Turnhout, K., & Terken, J. (2003). Facial orientation during multi-party interaction with information kiosks. In *Proceedings of the Interact 2003*.

Beskow, J., Edlund, J., Granström, B., Gustafson, J., Skantze, G., & Tobiasson, H. (submitted). The MonAMI Reminder: a spoken dialogue system for face-to-face interaction. Submitted to *Interspeech 2009*.

Horvitz, E., Kadie, C. M., Paek, T., & Hovel, D. (2003). Models of attention in computing and communication: from principles to applications. *Communications of the ACM*, 46(3), 52-59.

Katzenmaier, M., Stiefelhagen, R., Schultz, T., Rogina, I., & Waibel, A. (2004). Identifying the Addressee in Human-Human-Robot Interactions based on Head Pose and Speech. In *Proceedings of ICMI 2004*.

Oh, A., Fox, H., Van Kleek, M., Adler, A., Gajos, K., Morency, L-P., & Darrell, T. (2002). Evaluating Look-to-Talk: A Gaze-Aware Interface in a Collaborative Environment. In *Proceedings of CHI 2002*.