# Attention and Interaction Control in a Human-Human-Computer Dialogue Setting

**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

This paper presents a simple, yet effective model for managing attention and interaction control in multimodal spoken dialogue systems. The model allows the user to switch attention between the system and other humans, and the system to stop and resume speaking. An evaluation in a tutoring setting shows that the user's attention can be effectively monitored using head pose tracking, and that this is a more reliable method than using push-to-talk.

## 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 (Bohus & Horvitz, 2009). 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.

This paper presents a simple, yet effective model for managing attention and interaction control in multimodal (face-to-face) spoken dialogue systems, which avoids these simplifying assumptions. We also present an evaluation in a tutoring setting where we explore the use of head tracking for monitoring user attention, and compare it with a more traditional method: push-to-talk.

## 2 Monitoring user attention

In multi-party dialogue settings, gaze has been identified as an effective cue to help disambiguate the addressee of a spoken utterance (Vertegaal et al., 2001). When it comes to human-machine interaction, Maglio et al. (2000) showed that users tend to look at speech-controlled devices when talking to them, even if they do not have the manifestation of an embodied agent. Bakx et al. (2003) investigated the use of head pose for identifying the addressee in a multi-party interaction between two humans and an information kiosk. The results indicate that head pose should be combined with acoustic and linguistic features such as utterances length. Facial orientation in combination with speech-related features was investigated by Katzenmaier et al. (2004) in a human-human-robot interaction, confirming that a combination of cues was most effective. A common finding in these studies is that if a user does not look at the system while talking he is most likely not addressing it. However, when the user looks at the system while speaking, there is a considerable probability that she is actually addressing a bystander.

## 3 The MonAMI Reminder

This study is part of the 6th framework IP project MonAMI[1]. 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., 2009). The dialogue system uses Google Calendar as a backbone to answer questions about events. However,

[1] http://www.monami.info/

it can also take the initiative and give reminders to the user.

The MonAMI Reminder is based on the HIGGINS platform (Skantze, 2007). The 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). This is pretty much a standard dialogue system architecture, with some exceptions. Dialogue management is split into a Discourse Modeller and an Action Manager, which consults the discourse model and decides what to do next. There is also an Attention and Interaction Controller (AIC), which will be discussed next.

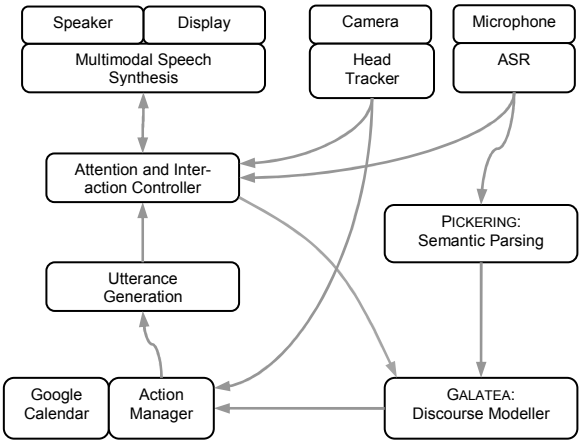


Figure 1. The system architecture in the MonAMI Reminder.

## 4 Attention and interaction model

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, shown in Figure 2. The states shown in the boxes can be regarded as the combined state of the system (columns) and the user (rows)[2]. Depending on the combined state, events from input and output components will have different effects. As can be seen in the figure, some combination of states cannot be realised, such as the system and user speaking at the same time (if the user speaks while the system is speaking, it will automatically change to the state INTERRUPTED). Of course, the user might speak while the system is speaking without the system detecting this, but the model should be regarded from the system's perspective, not from an observer.

The user's attention is monitored using a camera and an off-the-shelf head tracking software. As the user starts to look at the system, the state changes from NONATTENTIVE to ATTENTIVE. When the user starts to speak, a *UserStartSpeak* event from the ASR will trigger a change to the LISTENING state. The Action Manager might then trigger a *SystemResponse* event (together with what should be said), causing a change into the SPEAKING state. Now, if the user would look away while the system is speaking, the system would enter the HOLDING state – the system would pause and then resume when the user looks back. If the user starts to speak while the system is speaking, the controller will enter the INTERRUPTED state. The Action Manager might then either decide to answer the new request, resume speaking (e.g., if there was just a back-channel or the confidence was too low), or abort speaking (e.g., if the user told the system to shut up).

There is also a CALLING state, in which the system might try to grab the user's attention. This is very important for the current application when the system needs to remind the user about something.

### 4.1 Incremental multimodal speech synthesis

The speech synthesiser used must be capable of reporting the timestamp of each word in the synthesised string. These are two reasons for this. First, it must be possible to resume speaking after returning from the states INTERRUPTED and HOLDING. Second, the AIC is responsible for reporting what has actually been said by the system back to the Discourse Modeller for continuous self monitoring (there is a direct feedback loop as can be seen in Figure 1). This way, the Discourse Modeller may relate what the system says to what the user says on a high resolution time scale (which is necessary for handling phenomena such as backchannels, as discussed in Skantze & Schlangen, 2009).

Currently, the system may pause and resume speaking at any word boundary and there is no specific prosodic modelling of these events. The synthesis of interrupted speech is something that we will need to improve.

[2] This is somewhat similar to the "engagement state" used in Bohus & Horvitz (2009).

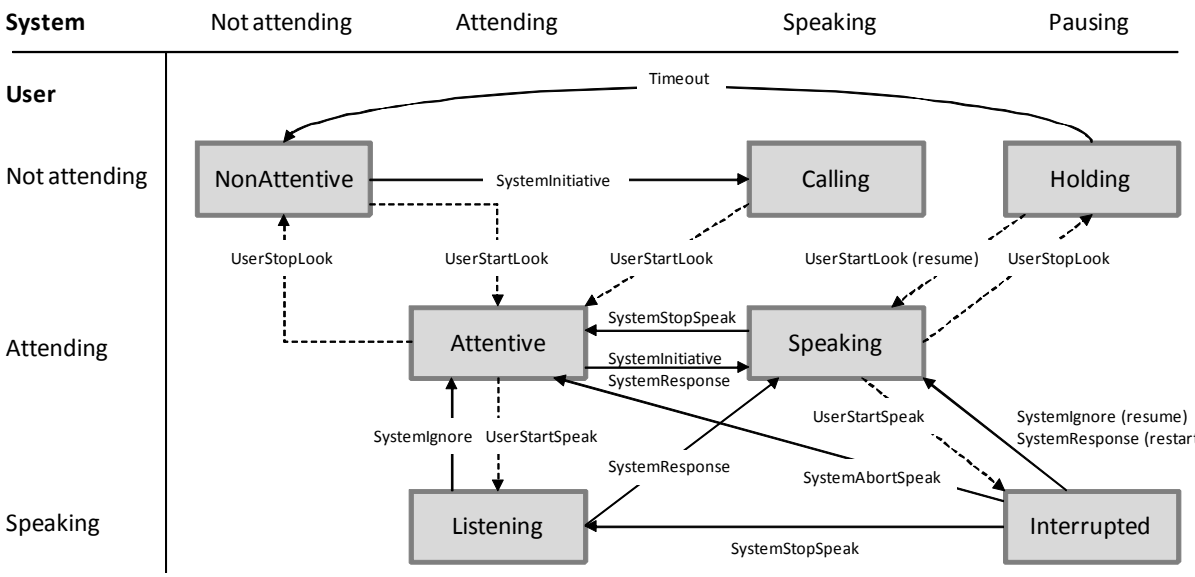


Figure 2. The attention and interaction model. Dashed lines indicate events coming from input modules. Solid lines indicate events from output modules. Note that some events and transitions are not shown in the figure.

An animated talking head is shown on a display, synchronised with the synthesised speech (Beskow, 2003). The head is making small continuous movements (recorded from real human head movements), giving it a more life-like appearance. The head pose and facial gestures are triggered by the different states and events in the AIC, as can be seen in Figure 3. Thus, when the user approaches the system and starts to look at it, the system will look up, giving a clear signal that it is now attending to the user and ready to listen.

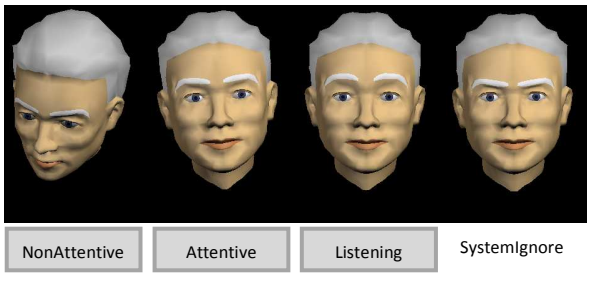


Figure 3. Examples of facial animations triggered by the different states and events shown in Figure 2.

## 5 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 4). 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). The tutor first explained both versions of the system to the subject and let her try both. The tutor gave the subjects hints on how to express themselves, but avoided to remind them about how to control the attention of the system, as this was what we wanted to test. After the introduction, the tutor gave the subject a task where both of them were supposed to find a suitable slot in their calendars to plan a dinner or lunch together. The tutor used a paper calendar, while the subject used the MonAMI Reminder. At the end of the experiment, the tutor interviewed the subject about her experience of using the system. 7 subjects (4 women and 3 men) were used in the evaluation, 3 lab members and 4 elderly persons in the target group (recruited by the Swedish Handicap Institute).

There was no clear consensus on which version of the system was the best. Most subjects liked the head tracking version better when it worked but were frustrated when the head tracker occasionally failed. They reported that a combined version would perhaps be the best, where head pose could be the main method for handling attention, but where a button or a verbal call for attention could be used as a fall-back.

When looking at the interaction from an objective point of view, however, the head tracking

Figure 4. 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

version was clearly more successful in terms of number of misdirected utterances. When talking to the system, the subjects always looked at the system in the head tracking condition and never forgot to activate it in the push-to-talk condition. However, on average 24.8% of all utterances addressed to the tutor in the push-to-talk condition were picked up by the system, since the user had forgotten to deactivate it. The number of utterances addressed to the tutor while looking at the system in the head tracking condition was significantly lower, only 5.1% on average (paired t-test; $p<0.05$).

These findings partly contradict findings from previous studies, where head pose has not been that successful as a sole indicator when the user is looking at the system, as discussed in section 2 above. 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.

Two of the elderly subjects had no previous computer experience. During pre-interviews they reported that they were intimidated by computers, and that they got nervous just thinking about having to operate them. However, after only a short tutorial session with the spoken interface, they were able to navigate through a computerized calendar in order to find two empty slots. We think that having a human tutor that guides the user through their first interactions with this kind of system is very important. One of the tutor's tasks is to explain why the system fails to understand out-of-vocabulary expressions. By doing this, the users' trust in the system is increased and they become less confused and frustrated. We are confident that monitoring and modelling the user's attention is a key component of spoken dialogue systems that are to be used in tutoring settings.

## 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. (2009). The MonAMI Reminder: a spoken dialogue system for face-to-face interaction. In *Proceedings of Interspeech 2009*.

Beskow, J. (2003). *Talking heads - Models and applications for multimodal speech synthesis*. Doctoral dissertation, KTH, Department of Speech, Music and Hearing, Stockholm, Sweden.

Bohus, D., & Horvitz, E. (2009). Open-World Dialog: Challenges, Directions, and Prototype. In *Proceedings of IJCAI'2009 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. Pasadena, CA.

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*.

Maglio, P. P., Matlock, T., Campbell, C. S., Zhai, S., & Smith, B. A. (2000). Gaze and speech in attentive user interfaces. In *Proceedings of ICMI 2000*.

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*.

Skantze, G., & Schlangen, D. (2009). Incremental dialogue processing in a micro-domain. In *Proceedings of EACL-09*. Athens, Greece.

Skantze, G. (2007). *Error Handling in Spoken Dialogue Systems – Managing Uncertainty, Grounding and Miscommunication*. Doctoral dissertation, KTH, Department of Speech, Music and Hearing, Stockholm, Sweden.

Vertegaal, R., Slagter, R., van der Veer, G., & Nijholt, A. (2001). Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of ACM Conf. on Human Factors in Computing Systems*.