The MonAMI Reminder: a spoken dialogue system for face-to-face interaction

Jonas Beskow[1], Jens Edlund[1], Björn Granström[1],
Joakim Gustafson[1], Gabriel Skantze[1] & Helena Tobiasson[2]

[1] KTH Speech Music & Hearing
[beskow,edlund,bjorn,jocke,gabriel]@speech.kth.se

[2] KTH Human-Computer Interaction Group
tobi@csc.kth.se

Abstract

We describe the MonAMI Reminder, a multimodal spoken dialogue system which can assist elderly and disabled people in organising and initiating their daily activities. Based on deep interviews with potential users, we have designed a calendar and reminder application which uses an innovative mix of an embodied conversational agent, digital pen and paper, and the web to meet the needs of those users as well as the current constraints of speech technology. We also explore the use of head pose tracking for interaction and attention control in human-computer face-to-face interaction.

1. Introduction

Information and communication technologies (ICT) play an increasing role in our lives, offering new services and facilitating communication between people. However, some of us are at risk of being excluded from these benefits. Two large and growing groups in this situation are elderly people, whose physical or mental functions may become reduced with age, and persons with permanent disabilities of all ages. They often find ICT services to be complicated, poorly designed, and not addressing their preferences and requirements. The goal of the 7th framework IP project MonAMI[1] is to address these problems – to develop and test new services based on existing technology, which are directly targeted towards and developed together with the elderly and disabled people who are going to use them. The project involves 4 FU centres (*Feasibility and Usability centres* where user tests are held in lab-like conditions) in 4 countries.

In this paper, we describe research on innovative interfaces in the MonAMI project. Specifically, we present the MonAMI Reminder, a multimodal spoken dialogue system designed to investigate hands-on two issues of direct relevance to the target group of the project: modality selection and ease-of-use. Our role in the project is to develop and test face-to-face interaction within the MonAMI context, taking a somewhat more long-term perspective, investigating the extent to which modern and innovative interfaces may improve MonAMI services, and what it takes to achieve such improvement. Our overall goal is to relieve human-computer interaction from some of the demands posed on the cognitive, visual and motor skills of the user, especially for elderly and disabled persons. We will evaluate conversational interfaces where the interaction metaphor [1] is shifted from desktop manipulation to face-to-face spoken dialogue with an embodied conversational agent (ECA). Furthermore, the domain is intimate and the potential users not necessarily at ease with technological aids, and in many cases, success may be a question of the users' trust and confidence in the good-will of the system. In this respect, the MonAMI reminder bears similarity to assistive and social spoken dialogue systems such as Wakamuru[2] and the Companion project [2].

Our first major target in this project is to demonstrate how a user-centred design approach can be used to *cure the pain* – how innovative interfaces can provide real solutions to real world problems. The second major research track is to explore the interactional abilities of the spoken dialogue system – to understand how a face-to-face setting may benefit users.

2. User-centered design

The MonAMI project focuses on real users who may be unfamiliar with recent technology, and is targeted at demonstrating and assessing accessible and affordable services for "people at risk of exclusion and loss of autonomy" (Mission statement, project overview), more specifically "elderly and disabled persons living at home". The target group is large, and can be expected to grow considerably: a high and increasing percentage of the population in the EU are of 80 years age or more.

The application to be developed in the project was chosen carefully in collaboration with our target users. A number of viable services were allocated for the Swedish FU centre in the project. With the help of in-depth interviews with two potential male users, we selected an advanced talking calendar – the MonAMI Reminder.

The two aforementioned persons had both had brain tumours recently and were thus suffering from cognitive disabilities. One of their major problems was to remember and initiate daily activities, ranging from taking a shower to meeting someone somewhere. Both used a range of applications and devices in order to organise their activities and be reminded about them: paper calendars, paper notes, PDA calendars, electronic whiteboards, and SMS notifications. Both interviewees felt comfortable using a paper calendar, but less so using electronic calendars. They did, however, have a strong need for the automatic notifications provided by the electronic solutions. Their current solutions involves duplicating events – most events go in the paper calendar, for easy browsing and editing, and the most important or easily forgotten events also go in one or more electronic devices to provide automatic reminders. The situation is further complicated by the involvement of care givers, who access some of the electronic devices but not all, and by the nature of the reminders. Examples are SMS text messages and PDA notifications, which are signalled with ring tones and/or flashing lights, which can make them obtrusive and difficult to tell from each other. Both persons expressed interest in using an ECA for getting notifications. We looked for a means of providing this while permitting users to keep their preferred method of organization, thus solving problems caused by its shortcomings.

In order to meet those requirements we designed a solution based on a mix of speech technology and a *digital pen and paper*. In this solution, the user can keep on using a paper calendar, but everything that is written is transferred to a cal-

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

[2] http://www.mhi.co.jp/kobe/wakamaru/english/

endar backbone using automatic handwriting recognition. The information in the calendar may then be accessed by the ECA so that the user can get notifications and ask questions about the content. Using the terminology McGee et. al. [3], a paper calendar is *augmented* with ECA technology. The result of such augmentation in a work practice is that neither the tools nor the way they are used are significantly changed. In our case this means that the users can continue using their paper calendar, but with the added possibility to ask about upcoming events and get oral notifications.

In order to transfer the ECA technology to the Swedish FU centre we later arranged a workshop about speech interface design. Members of the staff were introduced to the user-centered design methods, a first implementation of the MonAMI Reminder was demonstrated and a role playing session around speech control in this scenario was conducted. This workshop gave further insight into the needs of our target group. For example, we learned that users with Alzheimer need to be able to ask repeatedly about what they are going to do. Thus, the MonAmi Reminder would be very helpful in this situation, as caregivers often find these questions tiresome. Such users also have a need to be able to ask about past events, such as "When did I take my pills?" or "Have my children visited lately?" Old people with dementia also need encouragement to get started with their daily activities and support on how to carry them out.

2.1. Usage scenario

The MonAMI Reminder domain contains the following tasks: adding, leafing through them and editing events in a calendar in a familiar way (as this is something that already works well for many potential users, and in other cases is managed by care givers); asking about the contents of the calendar; and being reminded of events in the calendar in an efficient and unobtrusive manner. A usage scenario is presented in Table 1[1].

Table 1: *A usage scenario for the MonAmi Reminder.*

Monday 15.00	
Petra is visiting Stefan in his home.	
There is an ECA display mounted on the wall.	
1.Petra	I heard that they will show the movie Shadowlands at the theatre Astoria tomorrow at 7 o'clock.
2.Stefan	That would be great. I'll just check my calendar… It looks like I will do my laundry then, but I can move that to Wednesday. *(Stefan strikes out the laundry event, writes it in at Wednesday, and then writes in the new event).*
3.ECA	One event moved. One new event added.
4.Stefan	*(turns to the ECA)* Could you please remind me of the new event 1 hour ahead
5.ECA	I will remind you tomorrow at 6 o'clock.
Tuesday 13.0	
6.Stefan	*(turning to the ECA)* when will I meet Petra?
7.ECA	At seven o'clock you have written "See Shadowlands with Petra at Astoria". I will remind you 1 hour ahead.
8.Stefan	Ok.
Tuesday 18.00	
9.ECA	Stefan!
10.Stefan	Yes?
11.ECA	In one hour you have written "See Shadowlands with Petra at Astoria".
12.Stefan	Ok, you can remind me in 15 minutes again.

[1] A video showing a similar scenario can be seen at http://www.youtube.com/watch?v=G7J9_8PTw9w

The scenario illustrates some of the challenges for the MonAmi Reminder. First, the system must be able to recognise things that have been added to the calendar, as exemplified in turn 6. Second, the system must build a discourse model in order to be able talk about events as entities and refer to them using anaphoric expressions (turn 3-4 in the example). Third, the system must be able to converse in a multi-party setting where there may be other persons involved, so that it doesn't try to interpret for example utterance 2.

3. System architecture and components

The MonAMI Reminder is based on the Higgins platform [4], with a distributed architecture designed to cater to development and research needs: flexibility and ease-of-use. Higgins places few restrictions on components, which can be implemented in any language and may run on any platform. Components run asynchronously, possibly in separate processes, and communicate with XML-encoded messages.

The current configuration uses readily available, off-the-shelf components for all components not directly researched in the project. One reason for this is that it is easier to report, but the main reason is to show the multilingual properties of the system. Both the ASR and the synthesis we now use are available in a large number of European languages – an important consideration for a project with a practical bias. Interpretation and text generation need to be migrated from language to language. This is currently unavoidable since they use functionality that is currently unavailable in commercially available systems. These components and functionalities are also under investigation in the project.

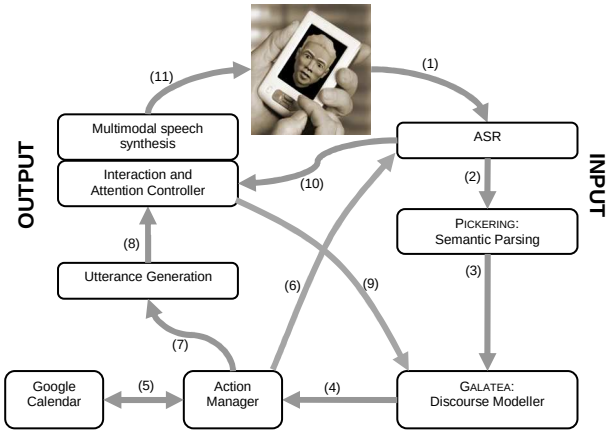

Figure 1: *The MonAmi Reminder architecture.*

Figure 1 shows the basic message flow between the components. In the current settings, we use a commercial ASR, as this enables porting to a large number of languages, as required by the project. The automatic speech recognition (ASR) is integrated through simple wrappers around the ASR API, providing access to voice activity detection (VAD), incremental results, on-the-fly grammar selection an word level confidence scores – all essential to other parts of the application.

The ASR passes the top hypothesis, with word confidence scores (2), to natural language understanding components: the robust interpreter *Pickering* [5], which makes a robust interpretation of this hypothesis and creates context-independent semantic representations of communicative acts (CAs), and the discourse modeller *Galatea* [6], for further interpretation taking dialogue context into account (3). The discourse model (4) is passed to an *Action Manager*, which initiates systems actions. *Google Calendar* is used as a backend for calendar information, and upon a request for information, the Action Manager searches Google Calendar (5) to generate a system

response in semantic form (7). This is passed to an utterance generation module which generates a textual representation that is forwarded to an *Interaction and Attention Controller* (8). This module controls the starting, stopping, pausing, resuming and monitoring of the speech output. The text-to-speech synthesis and facial animation (the Multimodal speech synthesis) is responsible for producing verbal as well as non-verbal responses from the system. The animated character is based on a 3D parameterised talking head that can be controlled by a text-to-speech system to provide accurate lip-synchronised audio-visual synthetic speech [7], and the commercially available *SynFace* [8], which comes in several languages, can be used as a replacement. As utterances are spoken by the system, the corresponding semantics and timings are sent back to Galatea for monitoring and inclusion in the discourse model (9).

4. Solutions under investigation

4.1. Unifying speech, pen and web

Using a spoken language interface in a calendar and reminder domain presents some hard speech technology challenges. The number of things that a person may want to be reminded about is almost indefinite, which is a problem for the ASR. Limited vocabularies are considerably easier for current ASR to work with than very large vocabularies. This conflict is addressed by mixing speech technology with the digital pen and paper. In this solution, the user can keep on using a paper calendar, but everything that is written is transferred to a calendar backbone. The information may then be accessed by the ECA so that the user can get notifications on what he has written in the calendar. The user may also ask questions such as "When was I supposed to meet Sara?" or "What's on my schedule today?"

The MonAMI Reminder is designed to combine and switch between several modalities – a pen and paper interface, speech, and a web based calendar – in a manner that meets all requirements gathered from the user interviews. Both interviewees thought that this was a very promising solution.

Pen and paper input and editing: Users are provided with a digital pen and a calendar made of special paper. To the user, the pen and calendar appears completely normal and they are used in exactly the same manner they are accustomed to. The pen captures entries and corrections written by the user and transfers them to a computer which performs handwriting recognition and passes the information to the calendar backbone. In the current version of the system, a commercial pen is used for this[1].

Google calendar perusal, input and editing: Regardless of how the MonAMI Reminder data is entered, it is stored in a Google Calendar. This makes it possible to browse and edit the calendar entries on the web for users who chose to do so. It also provides a uniform and easily accessible interface for care givers to edit and add entries. Naturally, as always with shared personal data, there are integrity issues involved. However, Google Calendar provides a fair set of tools to deal with this type of issue, such as the use of multiple calendars, some of which may be private (e.g. for the user's eyes only) and others shared over a group (e.g. care givers). Care must be taken, however, when adding, deleting or editing calendar entries through the web interface, since the paper calendar will not reflect these changes.

[1] Using the Anoto technology: http://www.anoto.com

4.2. Dynamic grammars

As stated above, using a static speech recognition language model would be impossible for this domain, as users are likely to ask for places and persons that will be out-of-vocabulary. Limited vocabularies are also considerably easier for current ASR to work with than very large vocabularies. In order to address this conflict we decided to use dynamic grammars that are updated based on the content of the user's calendar.

Each time the calendar gets updated, the Action Manager parses the events and extracts participants, event types and places. These are sent to ASR, so that the ASR grammars may be dynamically updated (path 6 in Figure 1). This makes it possible for the user to ask about events in the calendar. A set of grammar rules (which specifies what the user may say) are defined in the following way:

```
when am I meeting Person
when am I to be at Place
when do I have Event
when do I have Event DateTime
what happens DateTime
```

Words in title case are pointers to other rules. The `DateTime`-rule is defined in advance and matches date and time expressions, such as "April the third" or "three o'clock", while `Event`, `Place` and `Person` are dynamically defined. For example, if the user has written "lunch with Eva at the Ritz", a new expansion of the rule `Event` ("lunch with Eva at the Ritz") will be added, as well as the `EventType` ("lunch"), the `Person` ("Eva") and the `Place` ("the Ritz"). This makes it possible for the user to ask questions such as "when am I meeting Eva?" and "When will I have lunch?"

The user input to the system (the result from the ASR) is parsed using a set of parse rules that includes the same set of rules that were used to parse the events in the calendar. An example is shown in Figure 2. The example shows how the parse rules can mix pre-defined words with arbitrary strings, such as the name "Stina". The matched string can then be included in the resulting semantics.

The rules that match the sub-phrase "meeting with Stina" are the same as those used for parsing the calendar entries and building the calendar database. The database consists of a larger tree structure containing all events, as compiled from Google Calendar. The semantic tree structure built when parser user input is then used as a match expression (a sub-tree) by the Action Manager to search the database (as described in [4]).

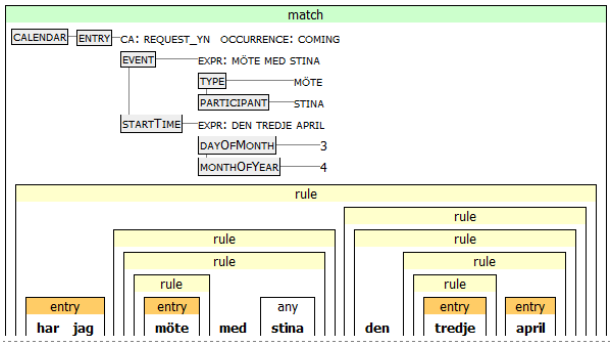

Figure 2: *Example parse of the sentence "Do I have a meeting with Stina on the third of April?" (in Swedish).*

Some notes on the relationship between the text input used for entries and the spoken dialogue are in order. Several potential spoken dialogue problems are alleviated by the text input: a) Parsing of the calendar allows us to represent the language models with regular grammars reflecting the current calendar content (to which dialogue state is added). The linguistic and contextual knowledge contained in these grammars improve recognition accuracy and disambiguate similar-sounding

words; b) Pronunciation modelling is done by loading a pre-compiled dictionary covering the vocabulary of the service. When users add new, unknown words and phrases using either of the text interfaces, they are given a pronunciation by a set of rules and added to the ASR vocabulary and the dynamic grammars, which minimizes occurrences of out-of-vocabulary words.

4.3. Utilizing face-to-face interaction

There are several aspects of spoken dialogue that are important if a spoken dialogue system is to be used by untrained users. Speech can solve the problems with obtrusive and confusing notifications and reminders, provided that the spoken dialogue system acts in a respectful manner. ECA technology promises to increase ease-of-use and to make the interaction more intuitive. One important aspect is to detect and display engagement behaviours which comprise both speech (including not just content, but timing of utterances) and non-verbal behaviour (including gaze and gesture), and are highly situated to the context of interaction and the actions of the other participants [9]. If an ECA displays engagement by producing believable and timely back-channel responses the users are more probable to be engaged in the interaction. Morency & Darrell [10] describe how head pose tracking was used in MACK and MEL as evidence for grounding and user engagement.

The MonAMI Reminder uses an ECA for notifications and permits the user to ask questions about upcoming events. The ECA may run on a wall-mounted display or on a PDA, or on several devices, so that the user may always choose the one that is most convenient. In a home environment it is important to model the user's focus of attention in order to know whether a speaking user is addressing the system or some other human being, as well as checking whether the user is paying attention to what is said by the system. Multimodal signals for turn-taking regulation can be expected to increase the robustness of the dialogue. There is well-documented relations between head pose and both attention and turn-taking – two areas we should attend to if we want to provide unobtrusive and respectful dialogue. Horvitz et. al. [11] used gaze and head pose tracking to decide whether a given utterance was directed at the computer in a command and control system and Bakx et. al. [12] used facial orientation to detect addressee in multi-party interactions with a information kiosks. The latter found that if the user was looking at a nearby person this was a reliable indicator that he was not addressing the system, while looking at the system was not a reliable indicator for addressing the computer. This phenomenon has also been observed by Ketzenmaier et. al. [13] who solves this by using a combination of acoustic and visual cues to determine addressee. We have previously demonstrated narrative ECAs that use head pose tracking to monitor listeners' attention and incremental speech synthesis to make it possible for the system to hold briefly when interrupted, then continue speaking in case the listener returns her attention, or cease speaking entirely if the listener remains inattentive [14]. As the prototype system we developed in the CHIL project has been very well received on several occasions (demonstrations at ICT, CHIL big meeting), we are investigating to what extent *look-to-talk* [15] may be used to control the dialogue. In the MonAmi system the behavior of the ECA is controlled by modeling the user's visual attention and spoken input. This model allows the ECA to:

- Only listen to what the user is saying while the user looks at the ECA.
- Provide attentional feedback when the user starts looking at the ECA.
- Only talk while the user is looking at the ECA and pause in the middle of utterances when the user looks away.
- Pause when user speech input is detected while the ECA is talking and possibly resume speaking, for example if the user provides verbal feedback or in the case of noise.
- Call for the user's attention, for example to remind the user of an upcoming event.

We are currently conducting user trials with the targeted user group at the Swedish FU centre, where we compare push-to-talk with look-to-talk for user attention control [16].

5. Acknowledgements

This research is carried out at KTH Speech Music & Hearing and the Centre for Speech Technology, a competence centre at KTH, supported by MonAMI, an Integrated Project under the European Commission's 6th Framework Program (IP-035147).

6. References

[1] Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech Communication, 50*(8-9), 630-645.

[2] Benyon, D., & Mival, O. (2008). Scenarios for Companions. In *Proceedings of Austrian Artificial Intelligence Workshop*.

[3] McGee, D., Cohen, P., & Wu, L. (2000). Something from nothing: Augmenting: a paper-based work practice with multimodal interaction. In *Designing Augmented Reality Environments Conference 2000* (pp. 71-80).

[4] Skantze, G. (2007). *Error Handling in Spoken Dialogue Systems - Managing Uncertainty, Grounding and Miscommunication*. Doctoral dissertation, KTH, Dept. of Speech, Music and Hearing.

[5] Skantze, G., & Edlund, J. (2004). Robust interpretation in the Higgins spoken dialogue system. In *Proceedings of ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*. Norwich, UK.

[6] Skantze, G. (2008). Galatea: A discourse modeller supporting concept-level error handling in spoken dialogue systems. In Dybkjær, L., & Minker, W. (Eds.), *Recent Trends in Discourse and Dialogue*. Springer.

[7] Beskow, J. (1997). Animation of talking agents. In Benoit, C., & Campbel, R. (Eds.), *Proc of ESCA Workshop on Audio-Visual Speech Processing* (pp. 149-152). Rhodes, Greece.

[8] Beskow, J., Karlsson, I., Kewley, J., & Salvi, G. (2004). SYNFACE - A talking head telephone for the hearing-impaired. In Miesenberger, K., Klaus, J., Zagler, W., & Burger, D. (Eds.), *Computers Helping People with Special Needs*, Springer-Verlag.

[9] Sidner, C. L., Kidd, C. D., Lee, C., & Lesh, N. (2004). Where to look: a study of human-robot engagement. In *Proceedings of the 9th international conference on Intelligent user interface* (pp. 78-84). New York, NY: ACM Press.

[10] Morency, L-P., & Darrell, T. (2004). From Conversational Tooltips to Grounded Discourse: Head Pose Tracking in Interactive Dialog Systems. In *Proceedings of the International Conference on Multi-modal Interfaces* (pp. 32-37).

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

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

[13] 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 International Conference on Multimodal Interfaces ICMI 2004*. PA, USA: State College.

[14] Beskow, J., Carlson, R., Edlund, J., Granström, B., Heldner, M., Hjalmarsson, A., & Skantze, G. (2009). Multimodal Interaction Control. In Waibel, A., & Stiefelhagen, R. (Eds.), *Computers in the Human Interaction Loop*. Berlin/Heidelberg: Springer.

[15] 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. *CHI 2002*.

[16] Skantze, G., & Gustafson, J. (submitted). Attention and interaction control in a human-human-computer dialogue setting. Submitted to *SigDial 2009*. London, UK.