Speech technology in the European project MonAMI

Jonas Beskow[1], Jens Edlund[1], Björn Granström[1], Joakim Gustafson[1], Oskar Jonsson[2] & Gabriel Skantze[1]
[1]Centre for Speech Technology, KTH Stockholm, Sweden
[2]Swedish Institute of Assistive Technology (SIAT), Vällingby, Sweden

Abstract

This paper describes the role of speech and speech technology in the European project MonAMI, which aims at "mainstreaming accessibility in consumer goods and services, using advanced technologies to ensure equal access, independent living and participation for all". It presents the Reminder, *a prototype embodied conversational agent (ECA) which helps users to plan activities and to remember what to do. The prototype merges speech technology with other, existing technologies: Google Calendar and a digital pen and paper. The solution allows users to continue using a paper calendar in the manner they are used to, whilst the ECA provides notifications on what has been written in the calendar. Users may also ask questions such as "When was I supposed to meet Sara?" or "What's on my schedule today?"*

Introduction

This paper presents the first version of a multimodal spoken dialogue system developed within the European project MonAMI (http://www.monami.info/). The objective of the MonAMI project is to demonstrate that accessible, useful services for elderly and disabled persons living at home can be delivered in mainstream systems and platforms. The technology platforms delivering the services are largely derived from standard technology, and integrate elements such as wearable devices, user interaction technology, and service infrastructures to ensure quality of service, reliability and privacy. The services are delivered on mainstream devices and services such as digital-TV, cell telephones and broadband Internet.

As traditional human-machine interfaces often assume a degree of computer literacy and are unintuitive to those unfamiliar with technology, development of innovative interfaces is also a part of the MonAMI project. The 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. Conversational interfaces are a radically different approach to human-machine interaction where the interaction metaphor is shifted from desktop manipulation to spoken dialogue, modelled on communication we are intrinsically familiar with: human-human face-to-face spoken dialogue. The result is an ECA – an *embodied conversational agent*, communicating with speech, facial expression, gaze and gesture. The innovative interfaces effort within MonAMI aims to develop interface technology based on the ECA; to implement a prototype that will be evaluated with users in the target group; and to adapt and use existing design and evaluation methods, based on end user involvement, for gaining understanding of IT functions and services that are considered meaningful by people with disabilities and people close to them. This demo paper presents the first version of *the Reminder*, the prototype ECA developed in the project in order to reach these goals.

The task

The choice of target application for an ECA prototype was informed by the services allocated for the Swedish FU centre (a *Feasibility and Usability centre* where user tests are held in lab-like conditions) in MonAMI, and in particular by meetings held with two potential users, both of whom have had a brain tumour and have cognitive disability, to identify potential areas addressing real key problems in their daily life. Based on these interviews, the choice fell on an application helping users plan activities and remember what to do. The overall application design is largely based on requirements from the interviews with the potential users, both of whom used a range of reminder applications and devices: paper calendars, paper notes, PDA calendars, electronic whiteboards, and SMS notifications, and both of whom expressed interest in using an ECA for getting notifications. The reminder addresses this by supporting pen input as well as speech, as seen in the following scenario:

December 14

10:00 *When speaking to Sara on the phone, Peter and Sara agree on a meeting at 12:00 the next day. Peter writes this down in his calendar.*

December 15

8:00 Peter: What happens today?
System: At 12 o'clock you have written "meeting with Sara".
Peter: Ok, remind me 1 hour ahead.

11:00 System: Peter?
Peter: Yes.
System: At 12 o'clock, you have written "meeting with Sara".
Peter: Ok.

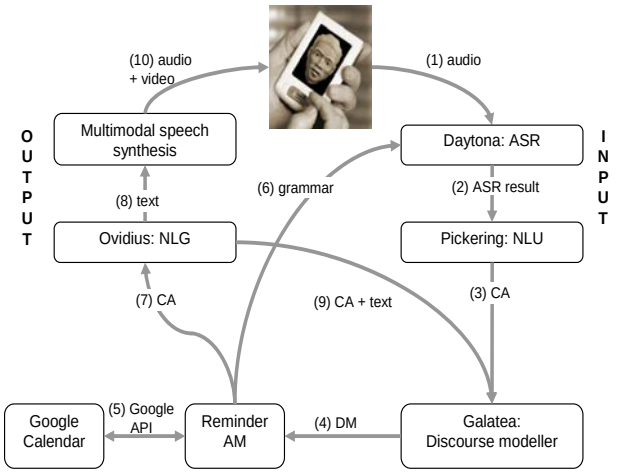

Figure 1: The Reminder architecture

The domain presents hard challenges for ECAs. For example, the things a person may want to be reminded of vary indefinitely, which is a problem for speech recognition.

The technology

The Reminder application architecture is based on the Higgins architecture (Edlund et al., 2004). The architecture is chiefly designed to cater to development and research needs, such as flexibility and ease of use, and places few constraints on components, which can be implemented in any language and run on any platform. Figure 1 shows the components and the message flow in the Reminder application.

From the ASR, the top hypothesis with word confidence scores (2) is forwarded for natural language understanding components. First it is sent to the robust interpreter Pickering (Skantze & Edlund, 2004), which makes a robust interpretation of this hypothesis and creates context-independent semantic representations of communicative acts (CAs). The results from Pickering (3) are forwarded to the discourse modeller Galatea (Skantze, in press), which may be regarded as a further interpretation step taking dialogue context into account. Galatea adds these to a discourse model (DM). The discourse model (4) is passed to the Reminder Action Manager, which initiates systems actions. The Reminder uses Google Calendar as its backend. When the discourse model is updated by a user request for calendar information, the action manager searches Google Calendar (5) to generate a system response in the form of a CA (7), which is passed to a component called Ovidius (Skantze, 2007). Ovidius generates a textual representation of the system utterance (8) that forwarded to a multimodal speech synthesiser for rendering (10). (The CA and the textual representation are both passed to Galatea (9) for inclusion in the discourse model.) The text-to-speech synthesis and facial animation is responsible of 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 (Beskow, 1997). The facial model includes control parameters for articulatory gestures as well as non-verbal expressions, which can be derived from motion recordings or developed using an interactive parameter editor (see Beskow et al., 2005 for details).

Each time the system initiates or if the Google Calendar entries are updated, the action manager also parses the calendar entries to build new speech recognition grammars and send them to the ASR (6). A schematic of Google Calendar can be seen in Figure 2, in which the original service interfaces are represented by solid lines and the extensions implemented in MonAMI by dotted lines. Utilising Google Calendar brings the obvious advantage of not having to provide hardware, software and connectivity for the calendar backbone, but there are several other benefits as well. Some of the more noteworthy come from the fact that the Google Calendar already provides user APIs in the form of a Web GUI for input and

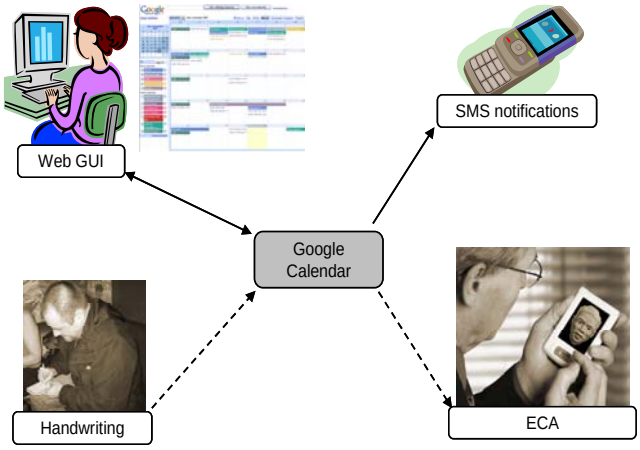

Figure 2: Calendar interfaces. Dotted lines added interfaces.

output, and SMS notifications as a form of output.

Mainly to meet the requirements from potential users, and partly in order to address the large and unknown vocabulary problem, we designed a solution based on a mix of speech technology and a *digital pen and paper*. To the user, the effect of the pen input is that of writing down events in a seemingly ordinary paper calendar. The text written by the user is transferred to a computer which performs handwriting recognition and transfers the information to a calendar backbone. The information may then be accessed by the ECA. The solution allows the users to use a paper calendar like they are used to, and addresses the ASR vocabulary problem: users may write anything they like in the calendar, but vocabulary can be limited to a base vocabulary the contents of calendar entries, which is used to update the vocabulary so that the user may speak about events in the calendar.

Next steps

The KTH Reminder is currently being prepared for a first set of evaluation experiments with potential users at SIAT (Swedish Institute of Assistive Technology/Hjälpmedelsinstitutet).

Acknowledgements

This research was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by MonAMI, an Integrated Project under the European Commission's Sixth Framework Program (IP-035147).

References

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.

Beskow, J., Edlund, J., & Nordstrand, M. (2005). A model for multi-modal dialogue system output applied to an animated talking head. In Minker, W., Bühler, D., & Dybkjaer, L. (Eds.), *Spoken Multimodal Human-Computer Dialogue in Mobile Environments, Text, Speech and Language Technology* (pp. 93-113). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Edlund, J., Skantze, G., & Carlson, R. (2004). Higgins - a spoken dialogue system for investigating error handling techniques. In *Proceedings of the International Conference on Spoken Language Processing, ICSLP 04* (pp. 229-231). Jeju, Korea.

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

Skantze, G. (in press). Galatea: A discourse modeller supporting concept-level error handling in spoken dialogue systems. To be published in Dybkjær, L., & Minker, W. (Eds.), *Recent Trends in Discourse and Dialogue*. Springer.

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