The OLGA project: An animated talking agent in a dialogue system

Jonas Beskow[1], Kjell Elenius[1] & Scott Mc Glashan[2]
[1]Department of Speech, Music and Hearing, KTH, Stockholm
[2]Swedish Institute for Computer Science, Stockholm

Abstract

The object of the Olga project is to develop an interactive 3D animated talking agent. The final target could be the future, digital TV-set, where the Olga agent would guide naive users through various new services on the networks. The current application is consumer information about microwave ovens. Olga implicates the development of a system with components from many different fields: dialogue management, speech recognition, multimodal speech synthesis, graphics, animation, facilities for direct manipulation and database handling. To integrate all knowledge sources Olga is implemented with separate modules communicating with a central dialogue interaction manager. In this paper we mainly describe the talking animated agent and the dialogue manager. There is also a short description of the preliminary speech recogniser used in the project.

Introduction

As spoken dialogue systems for simple information services begin to move from the laboratory into the area of technology, research interest is increasing turning to the integration of spoken dialogue interfaces with other modalities such as graphical interfaces. Apart from the general advantages of allowing an alternative input and output modality, speech can compensate for some of the apparent limitations of a graphical interface. Advantages include increased speed of interaction, higher bandwidth (attention and attitude expressed through stress and prosody, etc.) and ability to describe objects not visually present. Conversely, the graphical interface can compensate for limitations of speech e.g. by making immediately visible the effects of actions upon objects and indicating through the display which objects are currently salient for the system.

By including an animated agent in the interface, several positive effects can be anticipated. The system will seem more anthropomorphic, which will make users more comfortable with the dialogue situation. The character can provide a link between the spoken and the visual information domains, being able to refer to graphical items in the interface, using gaze and pointing. Body language, facial expression and gaze can potentially be very useful communication channels in a spoken interface. Furthermore, proper lip-synchronised articulation will improve intelligibility of system utterances (Beskow et. al. 1997). To date, some efforts has been made at combining spoken dialogue with human-like animated characters. Cassel et. al. (1994) have developed a conversation simulation system, with two animated talking and gesturing agents, whose intonation and gestures are controlled by rules. Nagao and Takeuchi (1994) proposed a simple scheme for displaying certain facial expressions as a supplement to speech during a man-machine dialogue. The Waxholm project at the Department of Speech, Music and Hearing (Bertenstam et al, 1995), which in some aspects can be seen as a predecessor to Olga, uses a human-like face for the talking agent, and utilises for example eye-gaze to refer to various on-screen items such as timetables.

The Olga Project

In the Olga project, we have developed a multimodal system combining a dialogue interface with a graphical user interface, which provides consumer information about microwave ovens. The system is composed of four main components: a speech and language understanding component; a direct manipulation interface which provides graphical information and widgets for navigation; an animated talking agent; and a dialogue manager for co-ordinating interpretation and generation in these modalities.

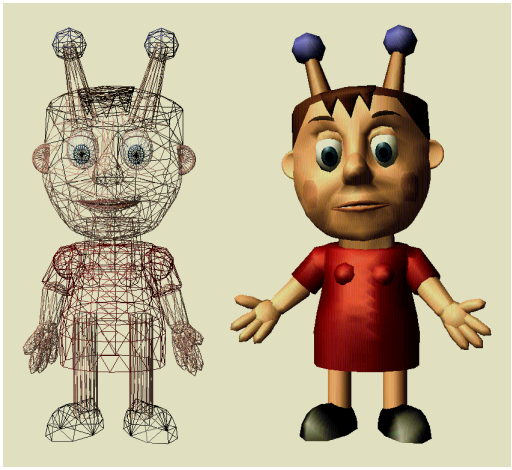

Figure 1: The Olga character

One of the motivating factors behind the Olga project was to ease the access to electronic information systems for people that are unfamiliar with computers. These people still constitute a substantial part of the population in all ages, with a predominance of elderly people. In the current implementation Olga deals with consumer information regarding microwave owens. One reason is that this application, though not very extravagant, indicates the ambition to make Olga an instrument for everyday help for people. Another reason is that the Swedish Consumer Agency (Konsumentverket) was participating during the initial stages of the project, and they had just finished a database with facts about microwave owens.

The Animated Agent

The Olga character is a three dimensional cartoon-like robot lady, that can be animated in real time. It is capable of text-to-speech synthesis with synchronised movements of lips, jaw and tongue. It also supports gesture and facial expression, that can be used to add emphasis to utterances, support dialogue turn-taking, visually refer to other on-screen graphics such as illustrations and tables, and to indicate the system's internal state: listening, understanding, uncertain, thinking (i.e. doing time-consuming operations such as searching a database) etc.

The Parameterised Polygon Model

The Olga character is implemented as a polygon model, consisting of about 2500 polygons, that can be animated at 25 frames per second on a graphics workstation. The character was first created as a static polygon representation of the body and head, including teeth and tongue. This static model was then parameterised, using a general deformation parameterisation scheme. The scheme allows a deformation to be defined by a few basic properties such as transformation type (rotation, scaling or translation), area of influence (a list of vertex-weight pairs that defines which polygon vertices should be affected by the transformation and to what extent) and various control points for normalisation of the deformation. It is then possible to define non-rigid deformations such as jaw opening, lip rounding etc, by combining basic deformations. Not only articulatory parameters, but also control of eyebrows, eyelids and smiling are defined in this manner. The body was parameterised by introduction of rotational joints at the neck, elbows, wrists, fingers etc.

Speech, Expression and Gesture

One important reason for using an animated agent in a spoken interface is that it actually will contribute, sometimes significantly, to the intelligibility of the speech, given that mouth movements are properly modelled (LeGoff et al., 1994). This is especially true if the acous-

tic environment is bad, due to for example noise or cross-talk, or if speech perception is impeded by hearing impairment. In a recent experiment, we found that the Olga-character increased the overall intelligibility of VCV-stimuli in noise from 30% for synthetic voice only, to 47% for the synthetic voice and synthetic face combination (Beskow et al., 1997).

Articulation is controlled by a rule-based text-to-speech system framework (Carlson and Granström, 1997). Trajectories for the articulatory parameters are calculated using a set of rules that account for coarticulation effects. This rule set was originally developed for an extended version of the Parke model (Parke, 1982), see Beskow (1995). However, since the parameters of the Olga model are chosen to conform to those of the extended Parke model, it is possible to drive Olga's articulation using the same set of rules. Once the parameter trajectories are calculated, the animation is carried out in synchrony with play-back of the speech waveform, which in turn is generated by a formant filter synthesiser controlled from the same rule-synthesis.

Non speech motion, such as gesture and facial expression, is controlled using a high-level scripting language (Tcl/Tk). Lists of time-value pairs for several parameters can be evaluated together in procedures. This allows for complex gestures, such as "shake head and shrug" or "point at graphics display", that require many parameters to be updated in parallel, to be triggered by a simple procedure call. Using procedures, it is easy to allow for whole gestures to be parameterised. For example, a procedure implementing a pointing gesture might take optional arguments defining direction of pointing, duration of the movement, degree of effort etc. During the course of the dialogue, appropriate gesture procedures are invoked in accordance to messages sent from the dialogue manager. There is also an "idle loop", invoking various gestures when nothing else is happening in the system.

The Dialogue Manager

The dialogue manager is based on techniques developed in a speech dialogue interface for telephone-based information systems in different languages (Eckert & McGlashan, 1993). A tri-partite model of interaction is responsible for semantic, task and dialogue interpretation. The semantics component provides a context-dependent interpretation of user input, and is capable of handling anaphora and ellipsis. A task component embodies navigation strategies to efficiently obtain information from the user necessary for successful database access. The dialogue component adopts an 'event-driven' technique for pragmatically interpreting user input, and producing system responses (Giachin & McGlashan, 1996). On the basis of user input events, it updates a dialogue model composed of system goals and dialogue strategies. The goals determine the behaviour of the system, allowing for confirmation and clarification of user input (to minimise dialogue breakdown), as well as requests for further information (to maximise dialogue progress). The dialogue strategies are dynamic so that the behaviour of the system varies with progress.

The dialogue manager needs to decide which modality to use for system output. In general, modality selection is defined in terms of characteristics of the output information, and the expressiveness and efficiency of the alternative modalities for realising it. Goals with a control or feedback function are realised in speech and gesture: for example, success in understanding user input is indicated with a head nodding gesture, while failure is indicated by speaking an explanation of the failure together with raised eyebrows and the month turned down. Product information is presented in speech and graphics: detailed product information is displayed while the agent gives a spoken overview.

The Speech Recogniser

The Olga project was originally planned for two years, and the addition of speech recognition was part of the second year. The intention was to make Wizard-of-Oz simulations during the first year in order to get speech and language material for the training of the recogniser. However, due to different circumstances it was evident that an Olga demonstrator had to be built during the first year, and in order to get a better impression of Olga's intended functionality, it was decided to include a preliminary speech recognition module based on the Waxholm recogniser (Ström, 1996). Some Olga specific features were introduced; the recogniser was modified to communicate with the dialogue interaction manager, and the Internet was used for inputting speech to the recogniser. The current version of the Olga speech recogniser is very preliminary and only able to recognise sentences according to the written scenario that forms the basis of the Olga demonstrator.

Acknowledgement

Olga involves a lot of interdisciplinary interaction. Besides the affiliations of the authors the Department of Computer Science at KTH and the Department of Linguistics at Stockholm University have participated. The Department of Linguistics at Helsinki University has developed a Swedish Constraint Grammar for syntactic analysis. Eva-Marie Wadman at Sweet was responsible for the artistic design of the graphical interface and the Olga-character. The project was initiated and managed by the Nordvis company. Olga has been funded by NUTEK, Stiftelsen för Kunskaps- och Kompetensutveckling, and Telia Research.

References

Bertenstam, J. Beskow, J., Blomberg, M., Carlson, R., Elenius, K, Granstrom, B., Gustafson, J., Hunnicutt, S., Hogberg, J., Lindell, R., Neovius, L., de Serpa-Leitao, A., Nord, L. and Ström, N. 1995. The Waxholm system - a progress report. *Proceedings of Spoken Dialogue Systems,* Vigsoe, Denmark.

Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E. and Öhman, T. 1997. The Teleface project - disability, feasibility and intelligibility. *In proceedings of Fonetik 97,* Umeå, Sweden.

Beskow, J. 1995. Rule-based Visual Speech Synthesis. *Proceedings of Eurospeech '95,* Madrid, Spain.

Carlson R. and Granström B. 1997. Speech Synthesis. *In* Hardcastle W. and Laver J. (eds) *The Handbook of Phonetic Sciences*. 768-788. Oxford: Blackwell Publishers Ltd.

Cassel, J., Steedman, M., Badler, N., Pelachaud, C., Stone, M., Douville, B., Prevost, S., Achorn, B. (1994), "Modeling the Interaction between Speech and Gesture", *Proceedings of 16th Annual Conference of the Cognitive Science Society*, Georgia Institute of Technology, Atlanta, USA.

Eckert, W. and McGlashan. S. 1993. Managing Spoken Dialogues for Information Services. *Proceedings of Eurospeech,* Berlin, Germany.

Giachin, E. and. McGlashan S. 1997. Spoken Language Dialogue Systems. *In* G. Bloothooft and S. Young (eds) *Corpus-based Methods in Language Processing.* Kluwer, The Netherlands.

Katashi Nagao and Akikazu Takeuchi. 1994. Speech Dialogue with Facial Displays: Multimodal Human-Computer Conversation. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, pp.102-109.

McGlashan, S. 1996. Towards Multimodal Dialogue Management. *Proceedings of Twente Workshop on Language Technology 11,* Enschede, The Netherlands.

Ström, N. 1996. Continuous Speech Recognition in the WAXHOLM Dialogue System. STL-QPSR, Department of Speech, Music and Hearing, KTH, 4/1996, 67-95.