

Dept. for Speech, Music and Hearing
**Quarterly Progress and
Status Report**

**OLGA - A dialogue system
with an animated talking
agent**

Beskow, J. and Elenius, K. O. E. and
McGlashan, S.

journal: TMH-QPSR
volume: 38
number: 2-3
year: 1997
pages: 001-006

<http://www.speech.kth.se/qpsr>



**KTH Computer Science
and Communication**

OLGA - A dialogue system with an animated talking agent*

Jonas Beskow¹, Kjell Elenius¹ & Scott McGlashan²

¹Department of Speech, Music and Hearing, KTH, Stockholm

²Swedish Institute for Computer Science, Stockholm

now at Ericsson Radio Systems, Stockholm

E-mail: beskow@speech.kth.se, kjell@speech.kth.se, scott.mcglashan@era-t.ericsson.se

Abstract

The object of the Olga project is to develop an interactive 3D animated talking agent. A futuristic application scenario is interactive digital TV, where the Olga agent would guide naive users through the various services available on the network. The current application is a consumer information service for microwave ovens. Olga required the development of a system with components from many different fields: multimodal interfaces, dialogue management, speech recognition, 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 com-

fortable 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, as shown in Beskow et al. (1997).

The Olga Project

In the Olga project, we have developed a multi-modal system combining a dialogue interface with a graphical user interface, which provides consumer information about microwave ovens.

The domain: Microwave ovens

An original motivation for the Olga project was to ease the access to electronic information systems for people who are unfamiliar with computers. They still constitute a substantial

* Also published in Proceedings of Eurospeech -97, 5th Eurospeech Conference on Speech Communication and Technology, Rhodes, Greece, Sept 22-25, 1997.

part of the population in all ages, with a predominance of elderly people. The selected consumer information application indicates the ambition to make Olga an instrument for the general public. Furthermore, the Swedish Consumer Agency (Konsumentverket) was participating during the initial stages of the project, and they provided a database with facts about microwave ovens.

Four main components

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.

Previous research

Compared with previous research in the area, the novelty of Olga lies in that it integrates interactive spoken dialogue, 3-D animated facial expressions, gestures, lip-synchronised audio-visual speech synthesis and a graphical direct manipulation interface. Cassel et al. (1994) have modelled speech and gesture in dialogue using two virtual agents, but no user interaction. Katashi & Akikazu (1994) employed animated facial expressions, but no gestures, as a back-channelling mechanism in a spoken dialogue system. Thorisson (1997) used a 2-D animated character together with input from many sources, including speech and gaze, to model mainly the social aspects of multimodal dialogue interaction. The Waxholm project at the Department of Speech, Music and Hearing (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 behaviour of the Olga agent is modelled using rules and parameterised templates; for example, the interaction strategies are based on condition-action rules where the condition part refers to the current interactional state as well as user input, and the action part to schematic descriptions of behaviour in language, graphics, and gestures. Similarly, in the animation module, realisation of a particular gesture is achieved by invoking the selected gesture's template and supplying appropriate parameters. This approach has worked well in our task domain allowing the agent's behaviour to be easily and quickly extended, as well as facilitating software maintenance.

The dialogue manager

The dialogue manager is based on techniques developed in a speech dialogue interface for telephone-based information systems in different languages (McGlashan, 1996, and Eckert & McGlashan, 1993).

A tri-partite model

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, compare Giachin & McGlashan (1997). 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. A more detailed description of the dialogue manager may be found in Beskow & McGlashan (1997).

Multimodality

In order to manage multimodal dialogues, input and output need to be informationally compatible at the dialogue management level. A user may provide input via buttons in the interface and the agent generate a spoken response; or a user may refer linguistically to an object which the agent realised graphically. Consequently, all input and output is represented in the semantic description language used for spoken input (McGlashan, 1996). This language also allows the user to use different modalities in the same response: e.g. clicking on an object, and then speaking a command to apply to it.

Output modality selection

The dialogue manager decides which modality to use for agent 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. In practice, selection is determined by rules which specify realisation in the three modalities depending on the action or state which the agent wants to express. Table 1 provides a simplified representation of rules (the rules can take in account other aspects of the action or state). 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 mouth turned down. In cases where the database access has required relaxation of product constraints, speech and a 'regret' gesture are realised. Product information itself is presented in speech and graphics: detailed product information is displayed while the agent gives a spoken overview. Finally, a print action is simply indicated with a graphical icon.

Table 1. Output modality selection rules.

Condition	Speech	Graphics	Gesture
reference success state	no	no	yes
reference failure state	yes	no	yes
constraint relaxation state	yes	no	yes
inform action	yes	yes	yes
printing action	no	yes	no

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 (Figure 1) 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 repre-

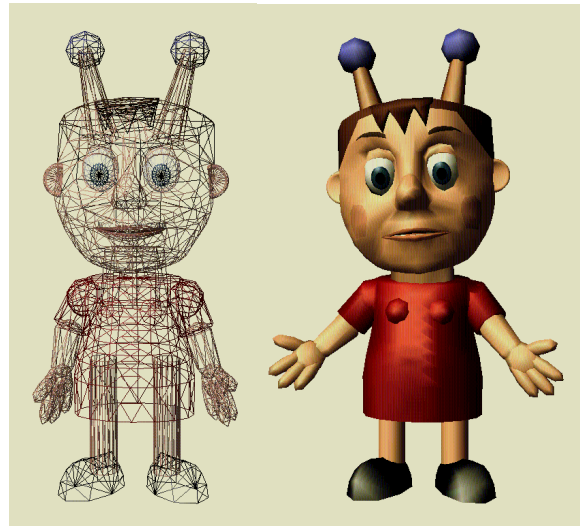


Figure 1. Wireframe and shaded representations of the Olga character.

sentation 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 and articulation

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, compare LeGoff et al. (1994). This is especially true if the acoustic 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 & Granström

(1997). Trajectories for the articulatory parameters are calculated using a set of rules that account for co-articulation effects. This rule set was originally developed for an extended version of the Parke model (1982) (Beskow, 1995). However, the articulation parameters of the Olga character are chosen to conform to those of the extended Parke model. This makes it 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 by the same rule-synthesis framework.

Complex gestures

Speech movements are calculated on an utterance-by-utterance basis and played back with high control over synchronisation. Body movements and non-speech facial expressions, on the other hand, place different requirements on the animation system. Say for example that we want the agent to dynamically change its expression during a user utterance, depending on the progress of the speech recognition. In this case, obviously utterance-by-utterance control will not do. The basic mechanism for handling this kind of movements in the Olga system is the possibility to, at any specific moment, specify a parameter trajectory as a list of time-value pairs to be evaluated immediately. Using such trajectory commands, gesture templates can be defined by grouping several commands together as procedures in a high-level scripting language (Tcl/Tk). 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.

Arguments for pointing and shrugging

Since a general scripting language is used, gesture templates can also be parameterised by supplying arguments to the procedure. For example, a pointing gesture might take optional arguments defining direction of pointing, duration of the movement, degree of effort, etc. As another example, a template defining a "shrug" gesture can have a parameter for selecting one from a set of alternative realisations, ranging from a simple eyebrow movement to a complex gesture involving arms, head, eyebrows and mouth corners. During the course of the

dialogue, appropriate gestures 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 scripting approach makes it easy to experiment with new gestures and control schemes. This sort of template based handling of facial expressions and gestures has proven to be a simple, yet quite powerful way of managing non-speech movements in the Olga system.

The Direct Manipulation Interface

All visualisation in the Olga system, except for the Olga agent, is controlled by the Direct Manipulation Interface, DMI. It manages graphics output as well as user initiated input. The output may be used for displaying interactive menus, information tables and/or visualisations, e.g., photos of specific microwave ovens. The graphics component of the DMI is based on the Distributed Interactive Virtual Environment developed at the Swedish Institute of Computer Science (Hagsand, 1996), which simplifies real-time manipulation of displayed 3D objects.

The speech recogniser

The Olga project was originally planned for two years, where the addition of the speech recogniser was scheduled for the second year. The intention was to make Wizard-of-Oz simulations with the Olga system during the first year of the project in order to collect speech and language material to be used for the training of the recogniser. However, due to various circumstances it later became evident that an Olga demonstrator had to be built during the first year. In order to get a better conception of Olga's intended functionality, it was decided to include a preliminary speech recognition facility.

The speech input module is based on the Waxholm recogniser described in Ström (1996). This is a software only continuous speech recognition engine with different modes for the phonetic pattern matching. In particular, standard multiple Gaussian mixtures and artificial neural networks are implemented for phone probability estimation. Thus, recognition may be performed either in a standard HMM or in a hybrid ANN/HMM framework. A lexicon with multiple pronunciations and a class bigram-grammar is used. The lexicon and grammar constraints are represented by a lexical graph, optimised for efficient lexical decoding. The

decoding is performed in a two-pass search. The first pass is a Viterbi beam-search and the second is an A* stack-decoding search. Multiple recognition hypotheses can be output either as standard N-best lists or in a more compact word-graph format.

The recogniser was modified to interact with the dialogue interaction manager and speech input was enabled over the Internet. 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

The original concept of Olga came from Henrik Wahlforss. He initiated and managed the project through the company Nordvis AB. The following organisations also took part: Centre for user oriented IT Design and Department of Speech, Music and Hearing, both at KTH; Department of Linguistics at Stockholm University and Swedish Institute of Computer Science. We would especially like to thank Eva-Marie Wadman and Olle Sundblad for graphical interface design and implementation. The Olga project was funded by Telia Research AB, NUTEK and Stiftelsen för kunskaps- och kompetensutveckling from September 1995 until September 1996.

References

- Beskow J, Dahlquist M, Granström B, Lundeberg M, Spens K-E, & Öhman T (1997). The Teleface project - Multimodal speech communication for the hearing impaired. In: *Proc of Eurospeech '97*, Rhodes, Greece.
- Cassel J, Steedman M, Badler N, Pelachaud C, Stone M, Douville B, Prevost S, Achorn B (1994). Modeling the interaction between speech and gesture. *Proc of 16th Annual Conference of the Cognitive Science Society*, Georgia Institute of Technology, Atlanta, USA.
- Katashi N & Akikazu T (1994). Speech dialogue with facial displays: multimodal human-computer conversation. *Proc of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, 102-109.
- Thórisson KR (1997). Gandalf: An embodied humanoid capable of real-time multimodal dialogue with people. *Proc of First ACM International Conference on Autonomous Agents*, Marina del Rey, California, pp 536-537.
- Bertenstam J, Beskow J, Blomberg M, Carlson R, Elenius K, Granström B, Gustafson J, Hunnicutt S, Högberg J, Lindell R, Neovius L, de Serpa-Leitao A, Nord L & Ström N (1995). The Waxholm system - a progress report. *Proc of Spoken Dialogue Systems*, Vigsoe, Denmark.
- McGlashan S (1996). Towards multimodal dialogue management. *Proc of Twente Workshop on Language Technology 11*, Enschede, The Netherlands.
- Eckert W & McGlashan S (1993). Managing spoken dialogues for information services. *Proc of Eurospeech '93*, Berlin, Germany, 3: 1653-1656.
- Giachin E & McGlashan S (1997). Spoken language dialogue systems. In: Bloothoof G & Young S, eds., *Corpus-based Methods in Language Processing*, Kluwer, The Netherlands.
- Beskow J & McGlashan S. Olga - a conversational agent with gestures. (To appear in *Proc of Workshop on Animated Interface Agents: Making them Intelligent*, Nagoya, Japan).
- Le Goff B, Guiard-Marigny T, Cohen M & Benoît C, (1994). Real-time analysis-synthesis and intelligibility of talking faces. *Proc of the Second ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, New York, USA.
- Carlson R & Granström B (1997). Speech Synthesis. In: Hardcastle W & Laver J, eds., *The Handbook of Phonetic Sciences*, Oxford: Blackwell Publishers Ltd, 768-788.
- Parke FI (1982). Parametrized models for facial animation. *IEEE Computer Graphics*, 2/9: 61-68.
- Beskow J (1995). Rule-based visual speech synthesis. *Proc of Eurospeech '95*, Madrid, Spain, 1: 299-302.
- Hagsand O (1996). Interactive MultiUser VEs in the DIVE System. *IEEE Multimedia Magazine*, vol 3, no 1.
- Ström N (1996). Continuous speech recognition in the WAXHOLM dialogue system. *TMH-QPSR, KTH*, Stockholm, Sweden, 4/1996: 67-95.