Experiences from the development of August - a multi-modal spoken dialogue system

Joakim Gustafson, Magnus Lundeberg and Johan Liljencrants

Centre for Speech Technology
Department of Speech, Music and Hearing, KTH
{joakim_g, magnusl, johan}@speech.kth.se

ABSTRACT

This paper describes experiences from the development of a Swedish spoken dialogue system with a talking agent, August. The system was exposed to the general public at a very early stage. Apart from describing the system and its components this paper discusses the problems encountered during the set-up and describe possible solutions considered. The system has been used for a period of six months to collect spontaneous speech data, largely from people with no previous experience of speech technology. One of the goals of the August project was to be able to analyze how novice users interact with a multi-modal information kiosk, placed without supervision in a public location. Another goal was to demonstrate how the speech technology modules developed at the department could be put together to rapidly prototype a multi-modal spoken dialogue system.

Keywords: multi-modal dialogue system, talking head, system development

1 INTRODUCTION

Speech technology promises to offer user-friendly interfaces for various information systems. Future dialogue systems will not only be used in laboratories by expert personnel, consequently they should be easy to use for people with little or no experience of computers. These systems might for example be set up as information kiosks in very diverse and technically difficult environments. A lot of questions need to be addressed in order to get these dialogue systems to work robustly in real-life applications, such as handling inexperienced users (sometimes with unrealistic expectations) and high background noise levels. These were some of the challenges of the August project. This paper gives an overview of the system and its different components. There is also a description of some of the problems encountered and the solutions considered.

2 THE AUGUST DIALOGUE SYSTEM

The August system was a Swedish multi-modal spoken dialogue system, featuring an animated agent (named after the 19th century author August Strindberg) with whom the user interacts. It was based on existing speech technology components developed at CTT and was built between January and August of 1998. Then the system was available daily for six months to any visitor at the Stockholm Cultural Centre, downtown Stockholm, as part of the *Cultural Capital of Europe '98* program. The users of the system were given very little or no information on how to interact with the system or what to expect. The animated agent communicated using synthetic speech, facial expressions and head movements [1]. In addition, August had a thought balloon in which additional textual information was displayed. The animated agent had a distinctive personality, which, as it turned out, invited users from the public to try the system and even socialize rather than just go for straightforward information-seeking tasks.

In order to elicit as spontaneous utterances as possible, the system was designed with a number of domains, instead of one single complex domain (such as e.g., ticket reservations). The simplest configuration of the August system presented information about restaurants and other facilities in Stockholm, about KTH, the research at CTT and about the system itself. August also had some basic knowledge about the life and works of August Strindberg. An important aspect of the project was that of handling multiple domains, even more work is needed to extend the existing domains and to add new ones.

The main goal of the August system was to study how naïve users would interact with a spoken dialog system covering several domains. In particular, it was interesting to study how users adapt their language when speaking to a computer. In the August system, the system responses differ both in length and complexity, from simple single-word utterances to long phrases with sub-clauses accentuated with both prosody and facial expressions. This resulted in a system that sometimes appeared to handle almost anything and generate very human-like dialogues, while it sometimes did not understand much at all. The data collect data was analyzed to see how users change their way of speaking during error resolution and what they said when the system responses were mostly adequate [2].

3 THE SYSTEM COMPONENTS

The system featured two computer screens, see Figure 1. One for the animated agent with a picture of Stockholm in the background and a thought balloon where information that was not synthesized could be displayed. A second screen was used for displaying textual database information as well as an interactive map for example used to show restaurants that matched the requirements of the user.

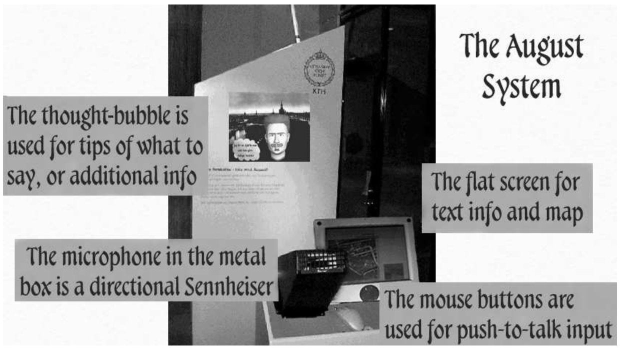

Figure 1. Some details in the set-up of the August system

The August system included the following components: A lip-synchronized 3D animated talking head, a camera eye which detected movement; a continuous speech recognizer; a system of simple dialogue managers; a semantic analyzer and a broker architecture for handling distributed modules. The communication between these modules is shown in Figure 2. The speech recognizer generated an n-best list of probable utterances as well as a confidence score. The n-best-list was sent to the semantic analyzer that extracted semantic information, such as domain, acceptability, and a set of semantic feature/value pairs. The domain prediction was used to determine which domain-specific dialogue manager to use. These dialogue managers worked independently to produce appropriate responses to send to the multi-modal synthesis module for generation. In some cases they also presented tables and maps on the other screen.

4 DEVELOPMENT ISSUES

A number of problems had to be addressed in the development of the August system. Four areas will be described in this paper: modularization and distribution of the speech technology components; speech input from inexperienced users in a noisy environment; designing a semantic analyzer of the input utterances that was easy to expand according to the user interaction; and finally developing a new talking head.

4.1 Modularization and a distributed architecture

The August system was developed from the speech technology components developed at KTH. The modules, such as the speech recognizer and the audio-visual speech synthesis, were provided with Tcl-interfaces. The Tcl/Tk language was chosen because it makes it easy to update the applications and since it has easy-to-use network capabilities. A broker architecture was also developed that handled the distributed system with servers and clients on several computers [3]. This was necessary since the animated agent operated on a Silicon Graphics, while all the other components were run on a Linux PC. All communication within the broker system was in text form to ensure portability and aid in debugging, while binary data, such as speech, was sent over separate TCP connections directly from producer to consumer. The Broker was written entirely in Java, and could therefore run on any modern computer. There were also tools that could display all communication between modules to facilitate debugging.

4.2 Audio input

The set-up environment for the August spoken dialogue system was tough in terms of acoustic conditions. It was a public space with a stone floor, glass walls, and background noise from other equipment and visitors. The simplest solution would have been to use a headset, but this was not feasible since the system was unsupervised. Instead a number of ways to mount a microphone out of reach from the users were considered. An initial idea was to use an acoustical lens in form of a large balloon, filled with CO_2. The speaker would then stand in front of the balloon and the microphone would be placed in the focal point at the other side. This did not work well because the sphere had poorly defined focal points and the balloon diameter of 1 m was too small to have appreciable effect at low frequencies. A second trial was to build a 1*1.3m segment of an ellipsoid reflector, where the focal points were located at the speaker position and 1 m above, respectively. Again, the basic

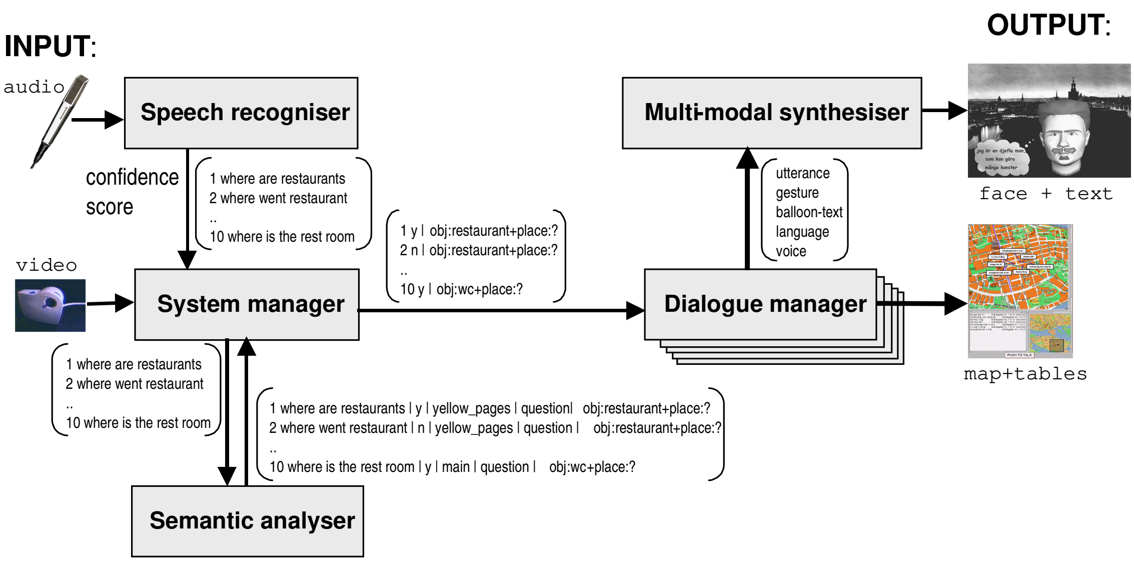

Figure 2. The components of the August system, and the information sent between them

problem was that the size of this reflector was too small to have an appreciable effect below about 1 kHz. Also a correction network was desirable to compensate for the reflector efficiency rising with frequency. The use of the reflector was found to be equivalent, in terms of speaker signal to background noise ratio, to speaking directly into the microphone from a distance of about 0.25 m. Getting sharper focussing would require a bigger reflector than was possible. The solution finally selected was to use a directional microphone, secured in a metal grid box, where the speaker could talk at short distance. The box introduced some deterioration of the sound but this did not effect the recognition significantly.

A number of directional microphones from Sennheiser were tested and compared to a conventional headset The following microphones were considered: **K6** - a back-electret condenser microphone including a head with a long gun pickup pattern; **MD421** - a dynamic microphone with a cardoid pick-up pattern and **MD441** - a dynamic microphone with super-cardoid pick-up pattern. To evaluate the usage of these microphones in our spoken dialogue system two subjects were recorded in the room where the August system was set-up. Twenty utterances included in the August system were recorded with the four microphones. Recognition results in Figure 3 indicate that the best performance was obtained using either the headset or the most expensive super-cardoid microphone, MD 441.

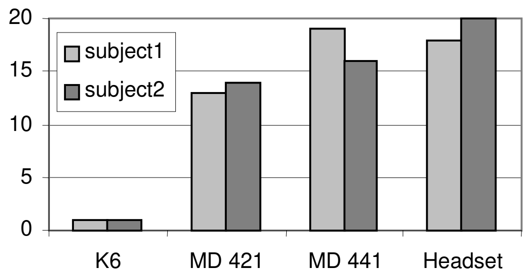

Figure 3. The number of recognized utterances for two subjects that uttered twenty August utterances

The long gun microphone could not be used at all, since it had to be placed at a distance from the users that was longer than feasible. The cheaper cardoid microphone picked up too much of the acoustic reflections and other room disturbances. Hence, we decided to use the MD441 microphone in the set-up of the August system.

4.3 Analyzing the user utterances

The system used a HMM-based recognizer with a main lexicon of about 500 words and idiomatic phrases. It generated an n-best list of utterance hypotheses, as well as a confidence score. The confidence score was computed by using two recognition engines in parallel: one with a lexicon of the words used in the system, and one that contained all permitted syllables in Swedish. Both engines generated an acoustic score for their outputs. A confidence score was obtained by subtracting the syllable score from the word score and normalizing the result by the utterance length. This gave a high score if the uttered string contained out-of-vocabulary words that had been forced to be recognized as words in the system lexicon. It gave a low score if the string was correctly recognized. This score was used in conjunction with the semantic analyzer described in the next section

One important topic in the August project was that of automating the process of extending the coverage of user utterances. The ultimate goal was to be able to extend an existing domain, or add a new one, with as little manual work as possible. The dialogue managers were kept very simple, since the complexity of the system was found in handling a number of simple domains instead of one complex. The dialogue managers could generate a number of possible answers to the semantic analysis of the user input. This was done by connecting a set of feature/value pairs to a number of pre-defined answers. The semantic analyzer that translated a recognized user utterance into the simple semantic representation had to be developed in a short time. The analyzer server was built around the freely available memory-based learning Timbl system [4]. An utterance hypothesis produced by the speech recognizer was given the analysis of similar examples in an annotated example database. A semantic analysis was obtained by simultaneously classifying an utterance along different dimensions. The semantic representation was shallow in that it consisted of a relatively simple feature-value structure, and was intended to make interesting distinctions from the dialogue system perspective rather than to constitute a "general" semantic component. There were three main fields that made up the semantic analysis, each of which was filled by an independent classifier. The first field stated whether an utterance was acceptable or not (*y* or *n*). (There is no clear-cut definition of what an unacceptable utterance is, but it was based on semantic grounds rather than grammatical ones only.) The second field predicted the domain of the utterance (e.g. *main, meta, strindberg, stockholm, yellow_pages…*) and the third field was instantiated with a flat feature-value representation of the utterance (e.g. *{object:restaurant, place:mariatorget}*).

The process of extending the coverage of user utterances to the August system involved the following steps. Firstly, a set of user utterances which the system could not handle, but should be able to handle, was collected. Secondly, each lexical item in the new utterances unknown to the system was provided with a phonetic transcription, a grammatical and a semantic tag and is then added to the lexicon. Thirdly, the new utterances were processed by the semantic analyzer, described above, and a human annotator corrected the analyses with the help of a simple graphical tool. If the new utterances had been recorded, they were processed by the speech recognizer, and the annotator corrected the semantic analysis of each unique hypothesis. The next time the system was started, the recognizer and the semantic analyzer were updated. However, in the initial state of the system, the dialogue manager which takes care of the semantic analyses, had to be manually updated to give a correct response to the extended repertoire of user input.

4.4 Developing a new talking head

A new lip-synchronized 3D talking head was developed for the project [1]. The purpose of developing a new face was to make use of experiences from previous projects and to create a unique character for the August system. In an earlier dialogue project, methods for adapting the audio-visual synthesis to new 3D-models were developed [5]. These methods have been improved and extended in the August project, in which the agent was made to look like the author August Strindberg. The purpose of creating a Strindberg lookalike was to show a well-known character; to indicate some knowledge about Stockholm, history and literature, and finally to give the agent a personality. Strindberg is famous for making some rather categorical statements about for example politics, women and reviewers.

When designing the agent, it was important that August should not only be able to generate convincing lip-synchronized speech, but also exhibit a rich and natural non-verbal behavior. To this end, a variety of gestures were developed. Among these gestures, six basic emotions were implemented to enable display of the agent's different moods. The synthetic speech output from August was also accentuated using non-articulatory head movements for example to accentuate focussed words. In early versions of the August system there was no immediate response from the system when a user asked the system a question. If the question resulted in a search on the Internet, users often perceived that the system did not receive the question, and therefore the user once again asked the same or a similar question. After finishing the search, the system would answer each question in order of appearance, resulting in a somewhat strange dialogue. To avoid this problem, and to enhance the perceived reactivity of the system, a set of listening gestures and thinking gestures was created. When the user pressed the push-to-talk button, the agent immediately displayed one out of ten listening gestures. At the release of the push-to-talk button, the agent changed to a randomly selected thinking gesture like frowning or looking upwards with the eyes searching, see Figure 4. The agent also used a desktop video camera together with image analysis software to be able detect the movements of the user. This made it possible to change the direction of the head and eyes to look at an approaching user [6].

Figure 4. The listening and thinking gestures of August

Speech synthesis parameter trajectories were generated by the KTH audio-visual text-to-speech system. Apart from generating the appropriate lip-movements in the animated face, these were also used as input to a Mbrola synthesizer for the sound generation [7]. The responses that were known in advance, including Strindberg quotations, were manually checked and changed.

5 CONCLUDING REMARKS

So far, the August system has been used by about 3000 people, which has generated a database of spontaneous man-machine interactions with the animated agent. The system was used by a diverse range of users in an acoustically hard environment. One of the aims of the project was to semi-automate the extension of the system according to the user interactions. Future work will include the development of more advanced domains. Work is also being done on allowing the dialogue manager to change the recognition lexicon and grammar depending on the dialogue

6 ACKNOWLEDGMENTS

We would like to thank Samsung for lending us the flat screen used in the system and Sennheiser for lending us a directional microphone. The August system used the Swedish male MBROLA-voice was created at the Dept. of Linguistics and Phonetics of Lund University. The following people also contributed to the development of the August system: Linda Bell, Jonas Beskow, Rolf Carlson, Björn Granström, Jesper Högberg, Erland Lewin, Nikolaj Lindberg, Kåre Sjölander, Eva-Lena Svensson and Tobias Öhman.

7 REFERENCES

1. Lundeberg, M. and Beskow, J. (1999) Developing a 3D-agent for the August dialogue system, To be Published in proceedings of AVSP'99.
2. Bell, L. and Gustafson, J. (1999) Utterance types in the August dialogues, To be published in Proceedings of IDS'99.
3. Lewin, E. (1997) The Broker Architecture at TMH, http://www.speech.kth.se/proj/broker/
4. Daelemans, W., Zavrel, J., van der Sloot, K. and van den Bosch, A. (1998) TiMBL: Tilburg Memory Based Learner, version 1.0, Reference Guide, *LK Technical Report 98-03*
5. Beskow, J. & McGlashan, S. (1997) Olga – A Conversational Agent with Gestures, In *Proceedings of the IJCAI'97 workshop on Animated Interface Agents - Making them Intelligent*, Nagoya, Japan, August 1997.
6. Öhman, T. (1999) A visual input module used in the August spoken dialogue system, to be published in QPSR 1-2/99
7. Dutoit, T., Pagel, V., Pierret, N., Bataille, F. and van der Vreken, O. (1996) The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes, In *Proc. of ICSLP'96, Philadelphia, vol. 3, pp. 1393-1396*