The August spoken dialogue system

Joakim Gustafson, Nikolaj Lindberg and Magnus Lundeberg

Centre for Speech Technology, KTH

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.

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].

The system featured two computer screens. 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. The August system was developed from the speech technology components developed at KTH. The 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 were as follows, 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.

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 either an acoustical lens in form of a large balloon, filled with CO_2 or a a 1*1.3m segment of an ellipsoid reflector. These solutions required too much space in order to work at lower frequencies. Instead we used 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

The system used a HMM-based recognizer with a main lexicon of about 600 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. 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}*).

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. The talking 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. 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. 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].

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. The responses that were known in advance, including Strindberg quotations, were manually checked and changed.

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

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) Dialogue Management in the August System, Proceedings of IDS'99.

3. [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*

4. Öhman, T. (1999) A visual input module used in the August spoken dialogue system, to be published in QPSR 1-2/99