THE AUGUST SPOKEN DIALOGUE SYSTEM

Joakim Gustafson, Nikolaj Lindberg and Magnus Lundeberg

Centre for Speech Technology '
Department of Speech, Music and Hearing, KTH
SE-100 44 Stockholm
{joakim_g, nikolaj, magnusl}@speech.kth.se

ABSTRACT

This paper describes the Swedish spoken dialogue system August. This system has been used to collect spontaneous speech data, largely from people with no previous experience of speech technology or computers. The aim was to be able to analyse how novice users interact with a multi-modal information kiosk, placed without supervision in a public location. The system described in this paper featured an animated talking agent, August. Speech data was collected during the six months that the system was exposed to the general public. The system and its components are briefly described, with references to more detailed papers.

Keywords: multi-modal dialogue system, talking head

1. INTRODUCTION

Future dialogue systems will not only be accessed through the telephone or via the Internet. Nor will they be used only in laboratories by expert personnel. The future systems should be easy to use for people with little or no experience of computers. Speech technology promises to offer user-friendly interfaces for dialogue systems. However, a lot of questions need to be addressed in order to get the dialogue systems to work robustly in real life applications. Future systems might be set up as information kiosks to be used in very diverse and technically difficult environments. Inexperienced users, possibly with unrealistic expectations as well as high background noise levels were some of the challenges of the August project. This paper gives an overview of the system and its different components. More detailed information about the individual components and the user interaction database collected during the project can be found in the references.

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. Between August 1998 and March 1999, it was available daily 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 users communicated with August by means of voice input only. The animated agent communicated using synthetic speech, facial expressions and head movements [1]. In addition, August had a thought balloon in which text that was not to be synthesised 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 socialise 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, though 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. Earlier studies of dialogue systems with a restricted domain and short system responses have shown that most users adapt their utterances by making them short and simple [2]. 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. The system responses were also pre-processed by manually adding prosodic information as well as face movements. 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 collected speech data can be analysed from two perspectives: Firstly, how do users react during error resolution, i.e., how they change their way of speaking when the dialogue fails [3, 4]. Secondly, what do they do in dialogues where the system responses are adequate in most dialogue turns, and the system appears reasonably intelligent. The speech data constitute an interesting database of user utterances, varying from simple greetings to more complex dialogues [5].

3. SYSTEM SET-UP

The August system was set-up in a public location with a lot of background noise from other equipment and visitors. Because of the tough acoustic conditions, a directional microphone, which was secured in a metal grid box, was used [6]. A push-to-talk button ensured that the system got only the utterances directed to the system and e.g. not to the company of the speaker. The speech technology components used to develop the August system were partly based on the software technology of our previous spoken dialogue systems [7, 8, 9]. The modules for speech recognition and audio-visual speech synthesis are provided with Tcl-interfaces, making it easy to change and extend the user interface and the functionality. A broker architecture that enabled a distributed system with servers and clients on several computers has also been developed [10]. Such a broker was necessary since the animated agent was implemented on a Silicon Graphics, while all the other components were on a Linux PC. The broker system relayed function calls between modules in text form over standard TCP internet connections.

The system featured two computer screens. The animated agent had its own screen, shown in Figure 1, with a picture of Stockholm in the background. This interface included a thought balloon where information that was not synthesised was displayed, such as tips on what to ask the system or additional information. A second screen was used for displaying textual and graphical database information, e.g. an interactive map showing restaurants and other facilities.

Figure 1. August quotes Strindberg: I am one hell of a man with a bag full of tricks.

4. THE SYSTEM COMPONENTS

The August system included the following components: A general purpose speech synthesis; a lip-synchronised 3D animated "talking head" with a rich repertoire of facial expressions; a camera eye which detects movement; a general purpose speech recogniser for continuous speech; different task domains; dialogue manager(s); example based semantic analyser (including topic prediction) and a general broker architecture for handling the distributed system modules, running under different platforms.

The user utterances were processed by the speech recogniser, which generated n-best lists of probable utterances. These were analysed to extract semantic information, such as domain, utterance type, 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 were supposed to work independently to produce appropriate responses to send to the multi-modal synthesis module for generation.

4.1. The recogniser

The system used a HMM-based speech recogniser [11] developed for the Waxholm dialogue system. The main lexicon of the test-system included only about 500 words (of which several were multi-word expressions) and a bigram class grammar of 70 classes and 229 class pairs. The speech recogniser generated an n-best list of the ten most probable 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 normalise 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 recognised as words in the system lexicon. It gave a low score if the string was correctly recognised. This score was used in conjunction with the semantic analyser described in the next section.

4.2. The utterance analyser

The semantic analyser component was one of the modules which was not already available, and had to be built when creating the August system. The "rapid prototyping" nature of the project put constraints on the semantic analyser: It had to be developed in a short time, and it should be robust and easy to extend. The semantic analyser component analysed the output from the speech recogniser, and translated a user utterance into a simple semantic representation, used by the dialogue manager to give the relevant response. It was fully example-based; no semantic or grammatical or rule based processing took place. The analyser server was built around the freely available memory-based learning Timbl system [12], developed at Tilburg University. An utterance hypothesis produced by the speech recogniser 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 and concatenating the results from the different sub-analyses. 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. For instance, number or gender agreement errors did not necessarily disqualify a hypothesis, while e.g., a semantically strange combination of verb and argument sometimes did so.) 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}*).

4.3. Dialogue manager

Since the complexity of the dialogue was found in handling a number of simple domains instead of one complex, the dialogue managers were kept very simple. They 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-processed answers. The dialogue manager of the Yellow pages-domain, however, used slot-filling in the generation of the responses.

4.4. Audio/visual speech synthesis

A new lip-synchronised 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 [8], methods for adapting the audio-visual synthesis to new 3D-models were developed. These methods have been improved and extended in the August project, in which the agent was made to look like the 19th century Swedish 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 (thus implying the domain of the system) and finally to give the agent a personality. Strindberg is famous for making some rather categorical statements about politics, women, reviewers, etc.

When designing the agent, it was important that August should not only be able to generate convincing lip-synchronised speech, but also exhibit a rich and natural non-verbal behaviour. 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. These emotions where made to resemble the six universal emotions described by Ekman [13].

Having the agent accentuate the auditory speech with non-articulatory movements is found to be very important with respect to the believability of the system. The main rules of thumb for creating the prosodic gestures have been to use a combination of head movements and eyebrow motion and to maintain a high level of variation between different utterances. A typical utterance from August could consist of either a raising of the eyebrow early in the sentence followed by a small vertical nod on a focal word or stressed syllable, or a small initial raising of the head followed by an eyebrow motion on selected stressed syllables. A small tilting of the head forward or backward often highlighted the ending of a phrase [17].

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, e.g. raising the eyebrows. 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 2.

Figure 2. The listening and thinking gestures of August.

In order to make the synthetic face appear less artificial, and to make the agent appear to be aware of the user's actions the agent changed the direction of the head and eyes according to the detected movements of an approaching user. This was accomplished by using a desktop video camera together with image analysis software tools [14, 15].

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 synthesiser for the sound generation [16]. The responses that were known in advance, including Strindberg quotations, were manually labelled with prosodic information. to further enhance the phrasing. A common problem was the distribution of stress, phrasing and prominence [18].

5. EXTENDING THE SYSTEM

One important topic, which has been at least partly addressed in the August project, is that of automating the process of extending the coverage of user utterances. The ultimate goal is to be able to extend an existing domain, or add a new one, with as little manual work as possible. Currently, the process of extending the coverage of user utterances to the August system involves the following steps. Firstly, a set of user utterances which the system cannot handle, but should be able to handle, is collected (either by introspection or, preferably, by analysing the database of genuine user utterances). Secondly, each lexical item in the new utterances unknown to the system is manually added to the lexicon. (A lexicon item has a phonetic transcription, a grammatical and a semantic tag.) Thirdly, the new utterances are processed by the semantic analyser, described above, and a human annotator corrects the analyses with the help of a simple graphical tool. If the new utterances have been recorded, they are processed by the speech recogniser, and the annotator corrects the semantic analysis of each unique hypothesis. The next time the system is started, the recogniser and the semantic analyser are updated. However, in the current state of the system, the dialogue manager which takes care of the semantic analyses, has to be manually updated to give a correct response to the extended repertoire of user input.

6. 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 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. The more than 12,000 sound files of user utterances will be further analysed. Work is being done on allowing the dialogue manager to change the recognition lexicon and grammar depending on the dialogue. Future work also includes experimenting with how the electronic eye could be used as input to the dialogue manager. This would make it possible to clear the dialogue history and generate instructions of how to use the system for approaching (silent) users.

7. ACKNOWLEDGMENTS

We would like to thank the friendly staff of Stockholms Akademiska Forum for allowing us to use their exhibition area for the August project. 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 created by Marcus Filipsson and Gösta Bruce of the Department of Linguistics and Phonetics of Lund University, Sweden.

The following people also contributed to the development of the August system: Linda Bell, Jonas Beskow, Rolf Carlson, Björn Granström, Jesper Högberg, Erland Lewin, Johan Liljencrants, Kåre Sjölander, Eva-Lena Svensson and Tobias Öhman.

8. 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, ESCA Workshop on Audio-Visual Speech Processing.

[2] Kennedy, A., Wilkes, A., Elder, L. & Murray, W. (1988) "Dialogue with machines", Cognition, 30 pp 73-105.

[3] Bell, L. and Gustafson, J. (1999), Repetition and its phonetic realizations: Investigating a Swedish database of spontaneous computer directed speech, to be published in Proceedings of ICPhS'99.

[4] Bell, L. and Gustafson, J. (1999) Interaction with an animated agent in a spoken dialogue system, Proceedings of Eurospeech '99.

[5] Bell, L and Gustafson, J (1999), Utterance types in the August dialogues, Proceedings of IDS'99, ESCA workshop on Interactive Dialogue in Multi-Modal Systems.

[6] Gustafson, J, Lundeberg, M and Liljencrants, J (1999), Experiences from the development of August - a multi-modal spoken dialogue system, Proceedings of IDS'99, ESCA workshop on Interactive Dialogue in Multi-Modal Systems.

[7] 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. and Ström, N. (1995): The Waxholm system – a progress report, In Proceedings of Workshop on Spoken Dialogue Systems, Vigsø, Denmark, May 1995

[8] 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.

[9] Sjölander, K., Beskow, J., Gustafson, J., Lewin, E., Carlson, R., and Granström, B. (1998) Web-based Educational Tools for Speech Technology, *proceedings of ICSLP9.*

[10] Lewin, E. (1997) The Broker Architecture at TMH, http://www.speech.kth.se/proj/broker/

[11] Ström, N. (1997), Automatic continuous speech recognition with rapid speaker adaptation for human/machine interaction, *PhD-Thesis Royal Institute of Technology, Stockholm.*

[12] 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*

[13] Ekman, P. (1979). About brows: Emotional and conversational signals. In. von Cranach, M., Foppa, K.,. Lepinies W. &. Ploog, D. (Eds.), Human ethology: Claims and limits of a new discipline: Contributions to the Colloquium (pp. 169-248). Cambridge: Cambridge University Press.

[14] Öhman, T.. (1998) The MODAL Video Module,

http://www.speech.kth.se/~tobias/MODAL/

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

[16] 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, *Proc. ICSLP'96, Philadelphia, vol. 3, pp. 1393-1396*

[17] Granström, B., House, D. and Lundeberg, M. (1999) Prosodic cues in multimodal speech perception, to be published in Proceedings of ICPhS'99.

[18] Svensson, E-L. (1999), Pronunciation in KTH:s text-to-speech system, Proceedings of Fonetik '99.