The Waxholm system - a progress report

Johan Bertenstam, Jonas Beskow, Mats Blomberg, Rolf Carlson, Kjell Elenius,
Björn Granström, Joakim Gustafson, Sheri Hunnicutt, Jesper Högberg, Roger Lindell,
Lennart Neovius, Lennart Nord, Antonio de Serpa-Leitao and Nikko Ström*

Department of Speech Communication and Music Acoustics,
KTH, Stockholm, Sweden

ABSTRACT

This paper describes ongoing development work on the spoken dialogue system, WAXHOLM, providing information on boat traffic in the Stockholm archipelago. The dialogue control and the natural language parser are implemented in an integrated, knowledge-based probabilistic language model. The recognition process is based on neural nets, A* lexical search, and a candidate reordering module. Speech synthesis for spoken response has been enhanced by the display of a synthetic, animated face. Application-specific data have been collected with the help of Wizard-of-Oz techniques.

1. INTRODUCTION

Our research group at KTH is currently building a generic system in which speech synthesis and speech recognition can be studied and developed in a man-machine dialogue framework. The system, previously reported in [1] and [2], is designed to facilitate the collection of speech and text data that are required for development.

The demonstrator application, which we call WAXHOLM, gives information on boat traffic in the Stockholm archipelago. It references time tables for a fleet of some twenty boats from the Waxholm company which connects about two hundred ports.

Besides the dialogue management and the speech recognition and synthesis components, the system contains modules that handle graphic information such as pictures, maps, charts, and time tables. This information can be presented as a result of the user initiated dialogue. The application has great similarities to the ATIS domain within the ARPA community, the Voyager system from MIT [3] and similar projects in Europe, for example SUNDIAL [4], the system for train timetables information developed by Philips [5], [6] and the Danish Dialogue Project [7].

The possibility of expanding the task in many directions is an advantage for our future research on spoken dialogue systems. In addition to boat time tables, the database also contains information about port locations, hotels, camping grounds, and restaurants in the Stockholm archipelago. This information is accessed with SQL, the standardized query language. An initial version of the system based on text input has been in operation since September 1992.

The system is implemented as a number of independent and specialised modules that run as servers on our computer system. A notation has been defined to control the information flow between them. The structure makes it possible to run the modules in parallel on different machines and simplifies the implementation and testing of alternate models within the same framework. The communication software is based on UNIX de facto standards, which will facilitate the reuse and portability of the components.

2. THE NATURAL LANGUAGE COMPONENT AND DIALOGUE MANAGEMENT

Our work on the natural language component is focused on a sublanguage grammar, a grammar limited to the particular subject domain -- that of requesting information from a travel database. Our parser, STINA, is knowledge based and is designed as a probabilistic language model. It contains a context-free grammar which is compiled into an augmented transition network (ATN). Probabilities are assigned to each arc after training. Characteristics of STINA are a stack-decoding search strategy, a feature-passing mechanism to implement unification and a robust parsing component.

Dialogue management based on grammar rules and lexical semantic features is implemented in STINA. The notation to describe the syntactic rules has been expanded to cover some of our special needs to model the dialogue. The STINA parser is running with two different time scales corresponding to the words in each utterance and to the turns in the dialogue. Topic selection is accomplished based on probabilities calculated from user initiatives.

We have found it very profitable to handle both the regular grammar analysis and the dialogue control with

*Names in alphabetic order

the STINA parser. The same notation, semantic feature system and developing tools can be shared. The rule-based probabilistic approach has made it reasonably easy to implement an experimental dialogue management module. STINA is described in more detail in [8], [9].

The parser has been evaluated in several different ways. Using about 1700 sentences in the Waxholm database as test material, 62 percent give a complete parse, whereas if we restrict the test data to utterances containing user initiatives (about 1200), the result is reduced to 48 percent. This can be explained by the fact that a large number of responses to system questions typically have a very simple syntax. If we exclude extralinguistic sounds such as lip smack, sigh and laughing in the test material based on dialogue initiatives by the user, the result is increased to 60 percent complete parses. Sentences with incomplete parses are handled by the robust parsing component and frequently effect the desired system response.

The perplexity on the Waxholm material is about 34 using a trained grammar. If extralinguistic sounds are taken away we get a reduction to about 30. If only utterances with complete parses are considered we get a perplexity of 23.

3. GRAPHICAL USER INTERFACE

The Waxholm system can be viewed as a micro-world, consisting of harbors with different facilities and with boats that you can take between them. The user gets graphic feedback in the form of tables complemented by speech synthesis. Up to now the subjects have been given a scenario with different numbers of subtasks to solve. A problem with this approach is that the subjects tend to use the same vocabulary as the text in the given scenario. We also observed that the user often did not get enough feedback to be able to decide if the system had the same interpretation of the dialogue as the user.

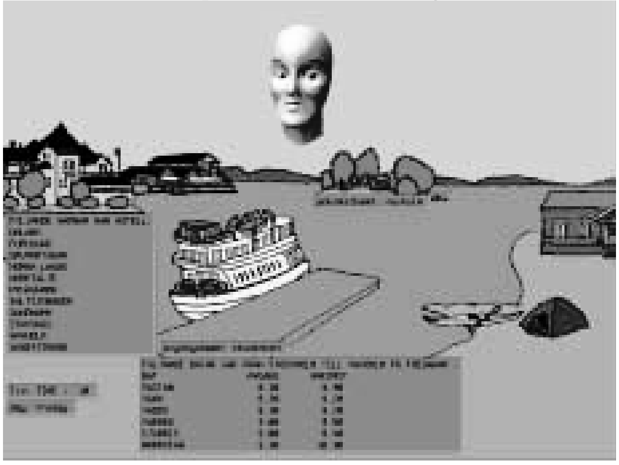

Figure 1. The graphical model of the WAXHOLM micro-world.

To deal with these problems a graphical representation that visualises the Waxholm micro-world is being implemented. An example is shown in Figure 1. One purpose of this is to give the subject an idea of what can be done with the system, without expressing it in words. Another purpose is that the interface continuously feeds back the information that the system has obtained from the parsing of the subject's utterance, such as time, departure port and so on. The interface is also meant to give a graphical view of the knowledge the subject has secured thus far, in the form of listings of hotels and so on.

4. SPEECH SYNTHESIS

For the speech output component we have chosen our multi-lingual text-to-speech system [10]. The system is modified for this application. The application vocabulary has been checked for correctness, especially considering the general problem of name pronunciation [11].

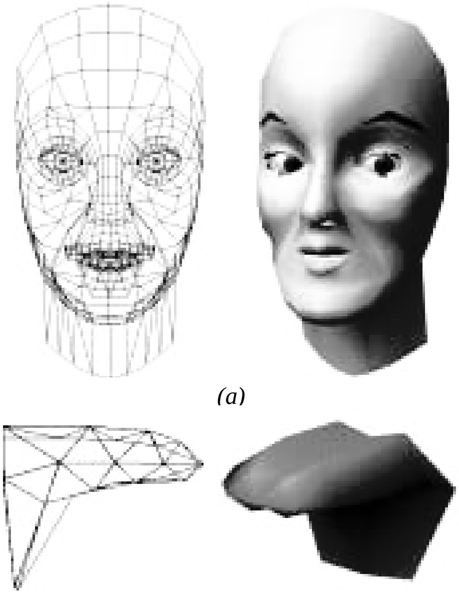

Figure 2. Wireframe and shaded representations of (a) the face and (b) the tongue models.

The speech synthesis has recently been complemented with a face-synthesis module (see Figure 2). Both the visual and the acoustic speech synthesis are controlled by the same synthesis software [12].

Since the recognition and synthesis modules have similar needs of semantic, syntactic and pragmatic information, the lexical information is shared. In dialogue applications such as Waxholm, we have a better base for prosodic modeling compared to ordinary text-to-speech, since, in such an environment, we will have access to much more information than if we used an unknown text as input to the speech synthesizer [13].

5. SPEECH RECOGNITION

The speech recognition component, which so far has only partially been integrated into the system, handles continuous speech with a vocabulary of about 1000 words. The work on recognition has been carried out along two main lines: artificial neural networks and a speech production oriented approach. Since neural nets are general classification tools, it is quite feasible to combine the two approaches.

Artificial neural networks

We have tested different types of artificial neural networks for performing acoustic-phonetic mapping of speech signals [14]. The tested strategies include self organising nets and nets using the error back propagation (BP) technique. The use of simple recurrent BP-networks has been shown to substantially improve performance. The self-organising nets learn faster than the BP-networks, but they are not as easily transformed to recurrent structures.

A* search

The frame-based outputs from the neural network form the input to the lexical search. There is one output for each of the 40 Swedish phonemes used in our lexicon. Each word in the lexicon is described on the phonetic level and may include alternative pronunciations of each word. The outputs are seen as the *a posteriori* probabilities of the respective phonemes in each frame. An A*, N-best search has been implemented using a simple bigram language model [15].

Candidate rescoring

The second step in the recognition process examines the output candidate list from the A* search. This search space is greatly reduced compared to the initial bigram model and a much more detailed analysis can be performed at this stage. Our system uses a formant-based speech production technique and a voice source model for the training of context-dependent phones [16]. One reason for this approach is the potential for reduction of the training and speaker adaptation data by utilising the close relation between phonemes in the production domain. Sharing of training data in parts of the production system is possible. For example, a small number of observations of voiced phonemes of an individual speaker can be used to adapt the voice source characteristic of the whole phoneme inventory. Phone duration information is also used in the evaluation process. For robustness reasons, the formant representation of the training data is transformed into the spectral domain for matching.

Despite the reduced search space, the reordering process still requires considerable processing time. For this reason, the acoustic rescoring of the candidates is performed after the recalculation of the linguistic scores by the STINA parser. The candidates are then remerged into a network, out of which only the best path is extracted.

Work is currently going on to integrate this component with the rest of the recognition module.

6. DATA COLLECTION

Speech and text data have been collected running the system with a Wizard of Oz replacing the speech recognition module [17]. The subjects are seated in an anechoic room in front of a display similar to Figure 1. The wizard is seated in an adjacent room facing two screens, one displaying what is shown to the subject and the other providing system information. All utterances are recorded and stored together with their respective label files. The label files contain orthographic, phonemic, phonetic and durational information. The phonetic labels derived from the input text are automatically aligned with the speech file [18], followed by manual correction.

All system information is logged during the data collection sessions making it possible to replay the dialogue. An experimental session starts with a system introduction, a reading of a sound calibration sentence and eight phonetically rich reference sentences. Each subject is provided with three information retrieval scenarios. Fourteen different scenarios have been used altogether, the first one being the same for all subjects.

So far 66 different subjects, of which 17 are female, have participated in the data collection sessions. The majority of the subjects, 43, were 20-29 years old while 4 were 30-39, 10 were 40-49 and 9 were more than 50 years old. Most subjects are department staff and undergraduate students from the school of electrical engineering and computer science. Some 200 scenarios have been recorded corresponding to 1900 dialogue utterances or 9200 words. The total recording time amounts to 2 hours and 16 minutes. There are more than 600 different words in the material but 200 suffice to cover 92% of the occurrences. The mean number of utterances for a scenario is about ten and the mean length of an utterance is five to six words. The dialogue is unrestricted in order to stimulate a natural interaction and to encourage user initiatives. However, subjects have proven very co-operative, with few exceptions, and answered system questions not using the opportunity to abruptly change the topic. Restarts are not as common as expected and can be found in less than 3% of the dialogue utterances.

In the label files, extralinguistic sounds are transcribed manually and labeled as interrupted words, inhalations, exhalations, clicks, laughter, lip smacks, hesitations and hawkings. Inhalations, often in combination with a smack, are the most common extralinguistic events. Inserted vowel sounds are also labeled. This kind of sound occurs when a consonant constriction is released. Vowel insertion and other extralinguistic sounds seem to be speaker specific features.

7. SUMMARY

The work on the Waxholm system is still in progress. The interactive development method, with Wizard of Oz simulations has given us a deeper understanding of the special needs in a spoken language dialogue system.

The unconstrained character of the task limits the performance of the speech recognition module. This is a challenge for further improvement of the system. Candidate reordering is expected to raise the recognition accuracy. Another possibility is to use dynamic language models, dependent on the previous dialogue.

Visual face synthesis complements the speech signal and is expected to raise the comprehension of spoken messages from the system.

The collected corpus contains a spectrum of different types of dialogue structure, speaking styles and speaker characteristics. Analysis of these data will help us model the dialogue and continue to improve the speech recognition performance.

8. ACKNOWLEDGMENT

This work has been supported by grants from The Swedish National Language Technology Program.

9. REFERENCES

[1] Blomberg, M., Carlson, R., Elenius, K, Granström, B., Gustafson, J., Hunnicutt, S., Lindell, R., Neovius, L. (1993), ‘An experimental dialogue system: WAXHOLM’, Proc. EUROSPEECH ‘93, Berlin, pp. 1867-1870.

[2] Carlson, R. (1994), ‘Recent developments in the experimental “Waxholm” dialog system’, ARPA Human Language Technology Workshop.

[3] Glass, J., Flammia, G., Goodine, D., Phillips, M., Polifroni, J., Sakai, S., Seneff, S., Zue V., ‘Multilingual spoken-language understanding in the MIT Voyager System’, Speech Communication (to be published).

[4] Peckham, J. (1993), ‘A new generation of spoken dialogue systems: results and lessons from the SUNDIAL project’, Proc. EUROSPEECH '93, Berlin.

[5] Aust, H., Oerder, M., Seide, F., Steinbiss V. (1994), ‘Experience with the Philips Automatic Train Timetable Information System’, Proc. IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA94), pp 67-72.

[6] Oerder, M., Aust, H. (1994), ‘A realtime prototype of an automatic inquiry system’, Proc. ICSLP 94, Yokohama, pp 703-706.

[7] Dalsgaard, P., Baekgaard, A. (1994), ‘Spoken language dialogue systems’, Proc. in Artificial Intelligence, Infix. Presented at the CRIM/FORWISS workshop on Progress and Prospects of Speech Research and Technology, Munich.

[8] Carlson, R., Hunnicutt, S. (1995), ‘The natural language component - STINA’, STL-QPSR 1/1995, Dept. of Speech Comm. and Music Acoustics, KTH, Stockholm (in print).

[9] Carlson, R., Hunnicutt, S., Gustafsson, J. (1995), ‘Dialog management in the Waxholm system’, ESCA/ETRW on Spoken Dialogue Systems, Vigsø, Denmark.

[10] Carlson, R., Granström, B., Hunnicutt, S. (1991), ‘Multilingual text-to-speech development and applications’, Advances in speech, hearing and language processing (Ainsworth AW, ed.), London: JAI Press, UK.

[11] Gustafson, J. (1994), ‘ONOMASTICA - Creating a multi-lingual dictionary of European names’, Working papers 43, Dept. of Linguistics and Phonetics, Lund University, Sweden.

[12] Beskow, J. (1995), ‘Rule-based visual speech synthesis’, Accepted for EUROSPEECH ‘95, Madrid.

[13] Bruce, G., Granström, B., Gustafson, K., Horne, M., House, D., Touati, P. (1995), ‘Towards an enhanced prosodic model adapted to dialogue applications’ ESCA/ETRW on Spoken Dialogue Systems, Vigsø, Denmark.

[14] Elenius K., Tråvén H. (1993), ‘Multi-layer perceptrons and probabilistic neural networks for phoneme recognition’, Proc. EUROSPEECH '93, Berlin.

[15] Ström, N. (1994), ‘Optimising the lexical representation to improve A* lexical search’, STL-QPSR 2-3/1994, Dept. of Speech Comm. and Music Acoustics, KTH, Stockholm, pp 113-124.

[16] Blomberg, M. (1994), ‘A common phone model representation for speech recognition and synthesis’, Proc. ICSLP 94, Yokohama, pp 1875-1878.

[17] Bertenstam, J., Blomberg, M., Carlson, C., Elenius, K., Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., Nord, L., Serpa-Leitao de, A., Ström, N. (1995), ‘Spoken dialogue data collected in the Waxholm project’, STL-QPSR 1/1995, Dept. of Speech Comm. and Music Acoustics, KTH, Stockholm.

[18] Blomberg, M., Carlson, R. (1993), ‘Labelling of speech given its text representation’, Proc. EUROSPEECH ‘93, Berlin, pp 1775-1778.