RECENT DEVELOPMENTS IN THE EXPERIMENTAL "WAXHOLM" DIALOG SYSTEM

Rolf Carlson

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

ABSTRACT

Recently we have begun to build the basic tools for a generic speech-dialog system. The main modules, their function and internal communication have been specified. The different components are connected through a computer network. A preliminary version of the system has been tested, using simplified versions of the modules. The dialog component of the system is described by a dialog grammar with the help of semantic features. Probabilities are also used in this process. We will give a general overview of the system and describe some of the components in more detail. Application-specific data are collected with the help of Wizard-of-Oz techniques. Currently the system is used during the data collection and the bionic wizard replaces only the speech-recognition module.

1. INTRODUCTION

Our research group at KTH* is currently building a generic system in which speech synthesis and speech recognition can be studied in a man-machine dialog framework. In addition, the system should facilitate the collection of speech and text data that are required for development. The system was first presented at the Eurospeech '93 conference [1]. The current paper is an expanded version of that paper. We will give a general overview of the system and describe some of the components in more detail. The dialog management component has recently been reformulated in a more general framework and is presented in the latter part of the paper.

2. THE DEMONSTRATOR APPLICATION

The demonstrator application, which we call WAXHOLM, gives information on boat traffic in the Stockholm archipelago (see Figure 1). It references time tables for a fleet of some twenty boats from the Waxholm company which connects about two hundred ports. Different days of the week have different time-tables.

Besides 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 to the user at his/her request. The application has great similarities to the ATIS domain within the ARPA community and other similar tasks in Europe, for example SUNDIAL. The possibility of expanding the task in many directions is an advantage for our future research on interactive dialog systems. An initial version of the system based on text input has been running since September 1992.

2.1. The database

In addition to boat time-tables the database also contains information about port locations, hotels, camping places, and restaurants in the Stockholm archipelago. This information is accessed by the standardized query language (SQL, Oracle). The time-table, which is the primary part of the database, brings some inherent difficulties to our application. One is that a boat can go in "loops," i.e. it uses the same port more than once for departure or arrival. This has been solved by giving unique tour identification numbers to different "loops." Another problem is that the port Waxholm may be used as a "transit port" for many destinations, and to avoid redundancy transit tours are not included in the database. Transits are instead handled by searching for tours from the departure port to Waxholm, and (backwards) from the destination port to Waxholm that require less than 20 minutes at the transit point [2].

2.2. Implementation

The dialog system is implemented as a number of independent and specialized modules that run as servers on our HP computer system. A notation has been defined to control the information flow between them. The structure makes it possible to run the system in parallel on different machines and facilitates 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.

* The Waxholm group consists of staff and students at the Department of Speech Communication and Music Acoustics, KTH. Most of the efforts are done part time. The members of the group in alphabetic order are: 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

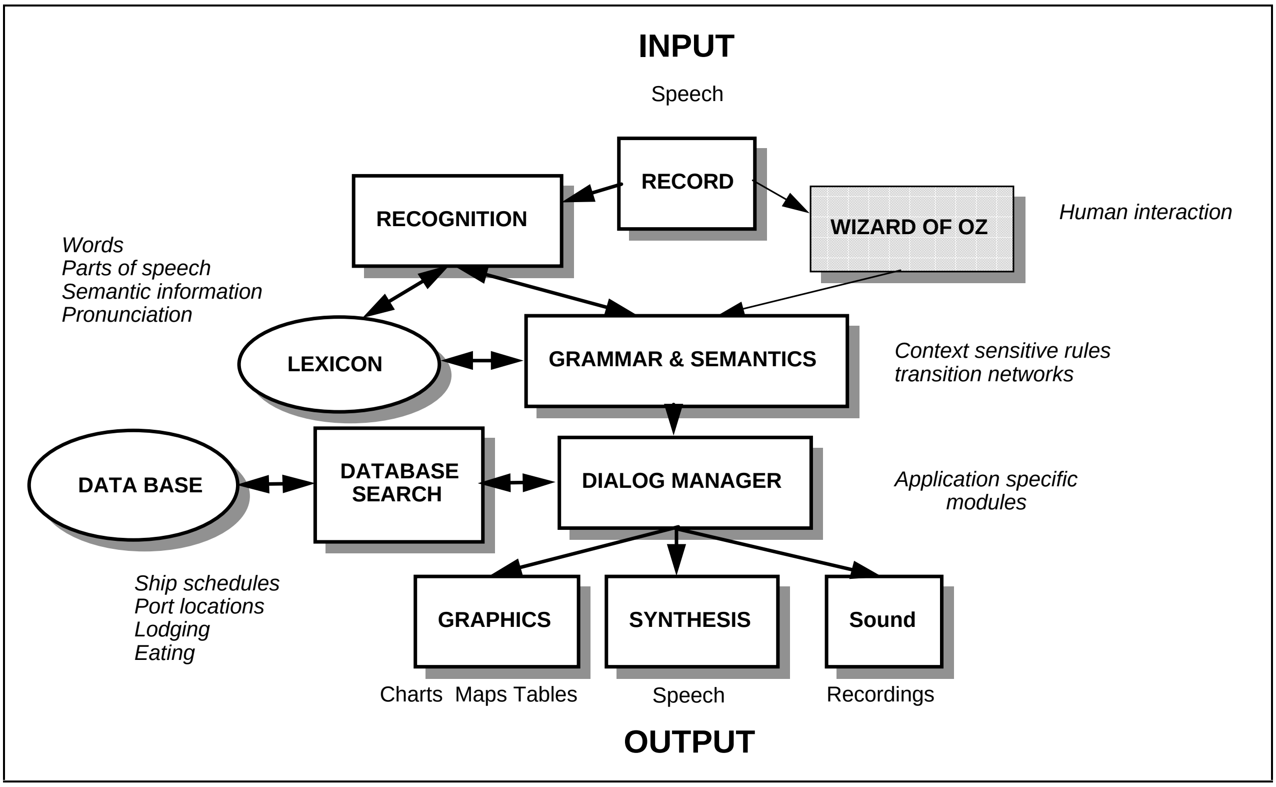

Figure 1. Block diagram of the demonstrator application Waxholm.

3. SPEECH RECOGNITION

The speech recognition component, which so far has not been integrated in the system during data collection, will handle 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.

3.1. Speech production approach

Our system uses a speech synthesis technique to generate spectral prototypes of words in a given vocabulary, see [3]. A speaker-independent recognition system has been built according to the speech production approach, using a formant-based speech production module including a voice source model. Whole word models are used to describe intra-word phonemes, while triphones (three-phoneme clusters) are used to model the phonemes at word boundaries. An important part of the system is a method of dynamic voice-source adaptation. The recognition errors have been significantly reduced by this method.

3.2. Artificial neural networks

We have tested different types of artificial neural networks for performing acoustic-phonetic mapping for speech signals, see [4], [5], and [6]. The tested strategies include self-organizing 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-organizing nets learn faster than the BP-networks, but they are not as easily transformed to recurrent structures.

3.3. Lexical 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. The lexicon may include alternate pronunciations of each word. The outputs are seen as the aposteriori probabilities of the respective phonemes in each frame. We have implemented an A* N-best search using a simple bigram language model. In a second stage the speech production approach mentioned above will be used to reorder the N-best list according to speaker specific criteria. A tight coupling between the parser and the recognizer is a long-term goal in the project. This will naturally influence the search algorithms.

4. SPEECH SYNTHESIS

For the speech-output component we have chosen the multi-lingual text-to-speech system developed in an earlier project [7]. The system is modified for this application. The application vocabulary must be checked for correctness, especially considering the general problem of name pronunciation.

Speaker-specific aspects are important for the acceptability of the synthetic speech. The WAXHOLM dialog system will focus our efforts on modelling the speaking style and speaker characteristics of one reference speaker. Since the recognition and synthesis modules have the same need of semantic, syntactic and pragmatic information, the lexical information will, to a great extent, be shared. The linguistic module, STINA, will also be used for improved phrase parsing, compared to the simple function-word based methods that have been used so far in the synthesis project. However, in dialog applications such as the proposed WAXHOLM demonstrator, information on phrasing and prosodic structure can be supplied by the application control software itself, rather than by a general module meant for text-to-speech. In a man-machine dialog situation we have a much better base for prosodic modelling compared to ordinary text-to-speech, since we, in such an environment, will have access to much more information than if we used an unknown text as input to the speech synthesizer.

5. NATURAL LANGUAGE COMPONENT

Our initial work on a natural language component is focused on a sublanguage grammar, a grammar limited to a particular subject domain: that of requesting information from a transportation database.

The fundamental concepts are inspired by TINA, a parser developed at MIT [8]. Our parser, STINA, i.e., Swedish TINA, is knowledge-based and is designed as a probabilistic language model [9]. It contains a context-free grammar which is compiled into an augmented transition network (ATN). Probabilities are assigned to each arc after training. Features of STINA are a stack-decoding search strategy and a feature-passing mechanism to implement unification.

In the implementation of the parser and the dialog management, we have stressed an interactive development environment. This makes it easier to have control over the system's progress as more components are added. It is possible to study the parsing and the dialog flow step by step when a tree is built. It is even possible to use the collected log files as scripts to repeat a collected dialog including all graphic displays and acoustic outputs.

5.1. Lexicon

The lexicon entries are generated by processing each word in the Two-Level Morphology (TWOL) lexical analyzer ([10] and [11]). Each entry is then corrected by removing all unknown homographs. New grammatical and semantic features, which are used by our algorithm and special application, are then added.

5.2. Features

The basic grammatical features can be positive, negative or unspecified. Unspecified features match both positive and negative features.

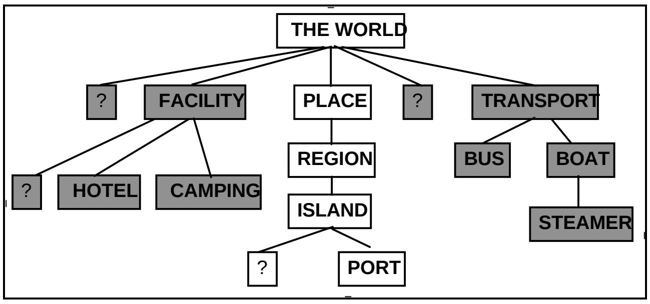

Figure 2. Example of a semantic tree feature structure.

Semantic features can be divided into two different classes. The basic features like BOAT and PORT give a simple description of the semantic property of a word. These features are hierarchically structured. Figure 2 gives an example of a semantic feature tree. During the unification process in STINA, all features which belong to the same branch are considered. Thus, a unification of the feature PLACE engage all semantic "non-shaded" features in Figure 2.

Another type of semantic feature controls which nodes can be used in the syntactic analysis. For example, the node DEPARTURE TIME cannot be used in connection with verbs that imply an arrival time. This is also a powerful method to control the analysis of responses to questions from the dialog module. The question "Where do you want to go?" conditions the parser to accept a simple port name as a possible response from the user.

6. DIALOG MANAGEMENT

6.1. Dialog rules

Dialog management based on grammar rules and lexical semantic features has recently been implemented in STINA. The notation to describe the syntactic rules has been expanded to cover some of our special needs to model the dialog. The STINA parser is running with two different time scales during data collection corresponding both to the words in each utterance and to the turns in the dialog. Syntactic nodes and dialog states are processed according to transition networks with probabilities on each arc.

Each dialog topic is explored according to the rules. These rules define which constraints have to be fulfilled and what action should be taken depending on the dialog history. Each dialog node is specified according to Figure 3.

Dialog Node Specifications

Node types
- branching or preterminal

Constraint evaluation on
- dialog flow features
- semantic frame slots and features
- ....

If more information needed
- synthesize question to user
- control parser to accept incomplete sentences
- ....

Node functions:
- record utterance
- synthesize message
- test constraints
- data base search using SQL
- graphic display table
- graphic display map
- graphic display picture
- ....

Figure 3. Dialog node specification.

The constraint evaluation is described in terms of features and the content in the semantic frame. If the frame needs to be expanded with additional information, a system question is synthesized. During recognition of a response to such a question the grammar is controlled with semantic features in order to allow incomplete sentences. If the response from the subject does not clarify the question, the robust parsing is temporarily disconnected so that specific information can be given to the user about syntactic or unknown word problems. At the same time a complete sentence is requested giving the dialog manager the possibility of evaluating whether the chosen topic is a bad choice.

A positive response from the constraint evaluation clears the way for the selected action to take place. The node function list in the figure gives examples of such actions.

6.2. Topic selection

In Figure 4 some of the major topics are listed. The decision about which path to follow in the dialog is based on several factors such as the dialog history and the content of the specific utterance. The utterance is coded in the form of a "semantic frame" with slots corresponding to both the grammatical analysis and the specific application. The structure of the semantic frame is automatically created based on the rule system.

TIME_TABLE
Goal: to get a time-table presented with departure and arrival times specified between two specific locations.
Example: När går båten? (When does the boat leave?)

GET_POSITION
Goal: to get a chart or a map displayed with the place of interest shown.
Example: Var ligger Vaxholm? (Where is Vaxholm?)

EXIST
Goal: to display the availability of lodging and dining possibilities.
Example: Var finns det vandrarhem? (Where are there hostels?)

OUT_OF_DOMAIN
Goal: Inform the user that the subject is out of the domain for the system.
Example: Kan jag boka rum. (Can I book a room?)

Figure 4. Some of the main topics used in the dialog.

Each semantic feature found in the syntactic and semantic analysis is considered in the form of a conditional probability to decide on the topic. The probability for each topic is expressed as: p(topic|F), where F is a feature vector including all semantic features used in the utterance. Thus, the BOAT feature can be a strong indication for the TIME-TABLE topic but this can be contradicted by a HOTEL feature.

6.3. Introduction of a new topic

The rule-based and to some extent probabilistic approach we are exploring makes the addition of new topics relatively easy. However, we do not know at this stage where the limits are for this approach. In this section we will give a simple example of how a new topic can be introduced.

Suppose we want to create a topic called "out of domain." Figure 5 illustrates the steps that need to be taken. First a topic node is introduced in the rule system. Some words will need to be included in the lexicon and labelled with a semantic feature showing that the system does not know how to deal with the subjects these words relate to. Then a synthesis node might be added with a text informing the user about the situation. Example sentences must be created that illustrate the problem. The dialog parser must be trained with these sentences labelled with the "out of domain" topic.

Since the topic selection is done by a probabilistic approach that needs application-specific training, data collection is of great importance for the progress of the project.

How to introduce a new topic

Introduce a new dialog grammar parent node

Expand the semantic feature set if needed

Specify dialog children nodes and their function and add to lexicon

Construct and label training sentences

Train topic probabilities

Figure 5. Introduction of a new topic.

The dialog will be naturally restricted by application-specific capabilities and the limited grammar. So far we also assume that the human subjects will be co-operative in pursuing the task. Recovery in case of human-machine "misunderstandings" will be aided by informative error messages generated upon the occurrence of lexical, parsing or retrieval errors. This technique has been shown to be useful in helping subjects to recover from an error through rephrasing of their last input [12].

7. DATA COLLECTION

We are currently collecting speech and text data using the WAXHOLM system. Initially, a "Wizard of Oz" (a human simulating part of a system) is replacing the speech recognition module, (See Figure 6). The user is placed in a sound-treated room in front of a terminal screen. The wizard sitting outside the room can observe the subject's screen on a separate display.

The user is initially requested to pronounce a number of sentences and digit sequences to practice talking to a computer. This material will be used for speaker adaptation experiments. After this the subject is presented with a task to be carried out. The scenario is presented both as text and as synthetic speech. An advantage of this procedure is that the subject becomes familiar with the synthetic speech. During the data collection, utterance-size speech files are stored together with the transcribed text entered by the wizard.

The stored speech files and their associated label files are processed by our text-to-speech system to generate a possible phonetic transciption. This transcription is then aligned and manually corrected. (For a description of this process see [13].)

The collected corpus is being used for grammar development, for training of probabilities in the language model in STINA, and also for generation of an application-dependent bigram model to be used by the recognizer. It is also being used to train word collocation probabilities. Our plan is to replace explicit formulations of semantic coupling by a collocation probability matrix.

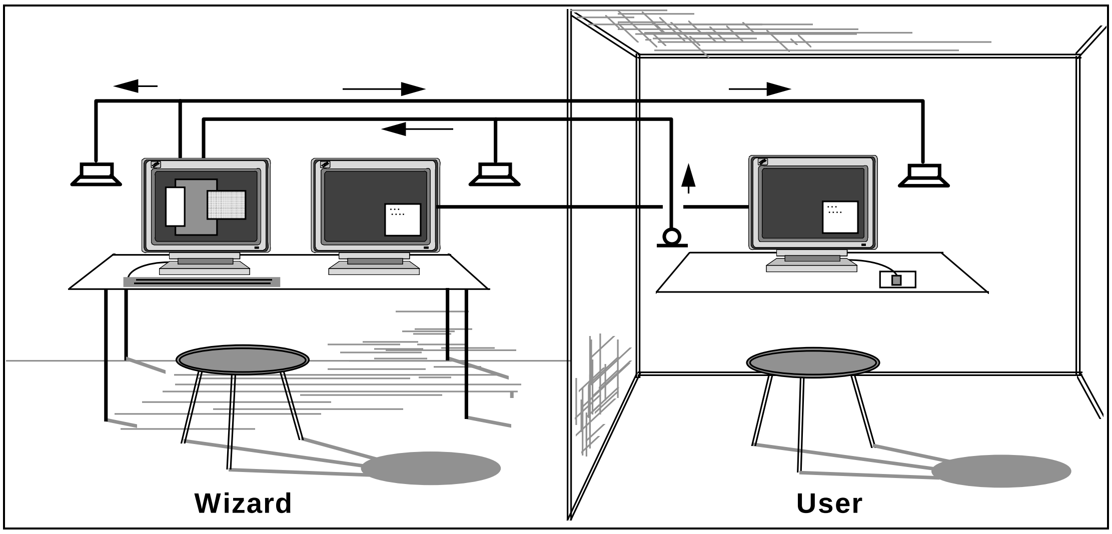

Figure 6. Hardware setup for data collection, with the help of a wizard.

8. FINAL REMARKS

In our presentation we have described the Waxhom project with special emphasis on the natural language components of the system. No module is yet considered complete. However, the most important work besides data collection is the integration of the speech recognizer into the system. The interaction between the parser and the recognizer still has to be improved.

The STINA parser has been expanded to better handle robust parsing and unknown word problems. In addition we are currently testing a simple application-independent grammar on unlimited text. This system will also be used as part of our general text-to-speech system, which is outside the scope of this presentation.

The dialog management module still needs to be tested in a more hostile environment. And the limits for our rule-based and probabilistic approach need to be explored.

ACKNOWLEDGEMENT

We thank Mike Phillips, SLS, MIT for valuable help in implementing the A*-search algorithm. This work has been supported by grants from The Swedish National Language Technology Program.

REFERENCES

[1] Blomberg, M., Carlson, R., Elenius, K., Granström, B., Gustafson, J., Hunnicutt, S., Lindell, R., and Neovius, L. (1993): "An experimental dialog system: WAXHOLM," Proceedings of Eurospeech '93. pp 1867-1870.

[2] Gustafson, J. (1992): "Databashantering som del av ett talförståelsesystem," Thesis work, Dept. of Speech Comm., KTH (only available in Swedish).

[3] Blomberg, M. (1991): "Adaptation to a speaker's voice in a speech recognition system based on synthetic phoneme references," Speech Communication, Vol. 10. pp 453-462.

[4] Elenius K. and Takács, G. (1990): "Acoustic-phonetic recognition of continuous speech by artificial neural networks," STL-QPSR 2-3, Technical Report, Dept. of Speech Comm., KTH, 1990. pp 1-44.

[5] Elenius, K. & Blomberg M., (1992): "Experiments with artificial neural networks for phoneme and word recognition," Proceedings of ICSLP 92, Banff, Vol. 2, pp. 1279-1282.

[6] Elenius K. & Tråvén H. (1993): "Multi-layer perceptrons and probabilistic neural networks for phoneme recognition," Proceedings of Eurospeech '93. pp 1237-1240.

[7] Carlson, R., Granström, B., & Hunnicutt, S. (1991), "Multilingual text-to-speech development and applications," (ed. A. W. Ainsworth), Advances in speech, hearing and language processing, JAI Press, London, UK.

[8] Seneff, S. (1989): "TINA: A Probabilistic Syntactic Parser for Speech Understanding Systems," Proceedings ICASSP-89, pp. 711-714.

[9] Carlson, R., & Hunnicutt, S. (1992): "STINA: A probabilistic parser for speech recognition," FONETIK'92, Sixth Swedish Phonetics Conference, May 20-22, 1992, Technical Report No. 10, Dept. of Information Theory, Chalmers University of Technology, Göteborg. pp 23-26.

[10] Koskenniemi, K. (1983): "Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production," University of Helsinki, Department of General Linguistics, Publications No. 11.

[11] Karlsson, F. (1990): "A Comprehensive Morphological Analyzer for Swedish," manuscript, University of Helsinki, Department of General Linguistics.

[12] Hunnicutt, S., Hirschman, L., Polifroni, J., & Seneff, S. (1992): "Analysis of the effectiveness of system error messages in a human-machine travel planning task," ICSLP 92 Proceedings, Vol. 1, University of Alberta, Canada. pp 197-200.

[13] Blomberg, M., & Carlson, R. (1993): "Labelling of speech given its text representation," Proceedings of Eurospeech '93. pp 1775-1778.