THE WAXHOLM SPOKEN DIALOGUE SYSTEM

ROLF CARLSON AND BJÖRN GRANSTRÖM*

Department of Speech Communication and Music Acoustics,
KTH, Stockholm, Sweden
*Names in alphabetic order

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 and an improved dialogue related prosody model. Application-specific data have been collected with the help of Wizard-of-Oz techniques.*

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 has been presented on several occasions, for example, the Eurospeech '93 conference (Blomberg et al., 1993), the ARPA meeting '94 (Carlson, 1994) and the ETRW on Spoken Dialogue Systems (Bertenstam et al., 1995a). It is designed to facilitate the collection of speech and text data that are required for development.

Spoken dialogue management has attracted considerable interest during the last years. Special workshops and symposia, for example the special workshop at Waseda University, Japan 1993 (Shirai and Furui, 1995), the AAAI 1995 Spring Symposium: Empirical methods in discourse interpretation and generation, Stanford University, USA and the 1995 ESCA workshop on Spoken Dialogue Systems in Vigsø, Denmark, have all been arranged to forward research in this field. We will not attempt to review this growing field in this paper. We will, however, describe in some detail our current effort to build a multi-modal dialogue system.

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 (Glass et al., 1994) and similar tasks in Europe, for example SUNDIAL (Peckham, 1993), the systems for train timetables information developed by Philips (Aust et al., 1994; Oerder and Aust, 1994) and CSELT (Clementino and Fissore, 1993; Gerbino and Danieli, 1993) and flight information in the Danish Dialogue Project (Dalsgaard and Baekgaard, 1994).

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 standardised 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. A block diagram of the system can be seen in Figure 1.

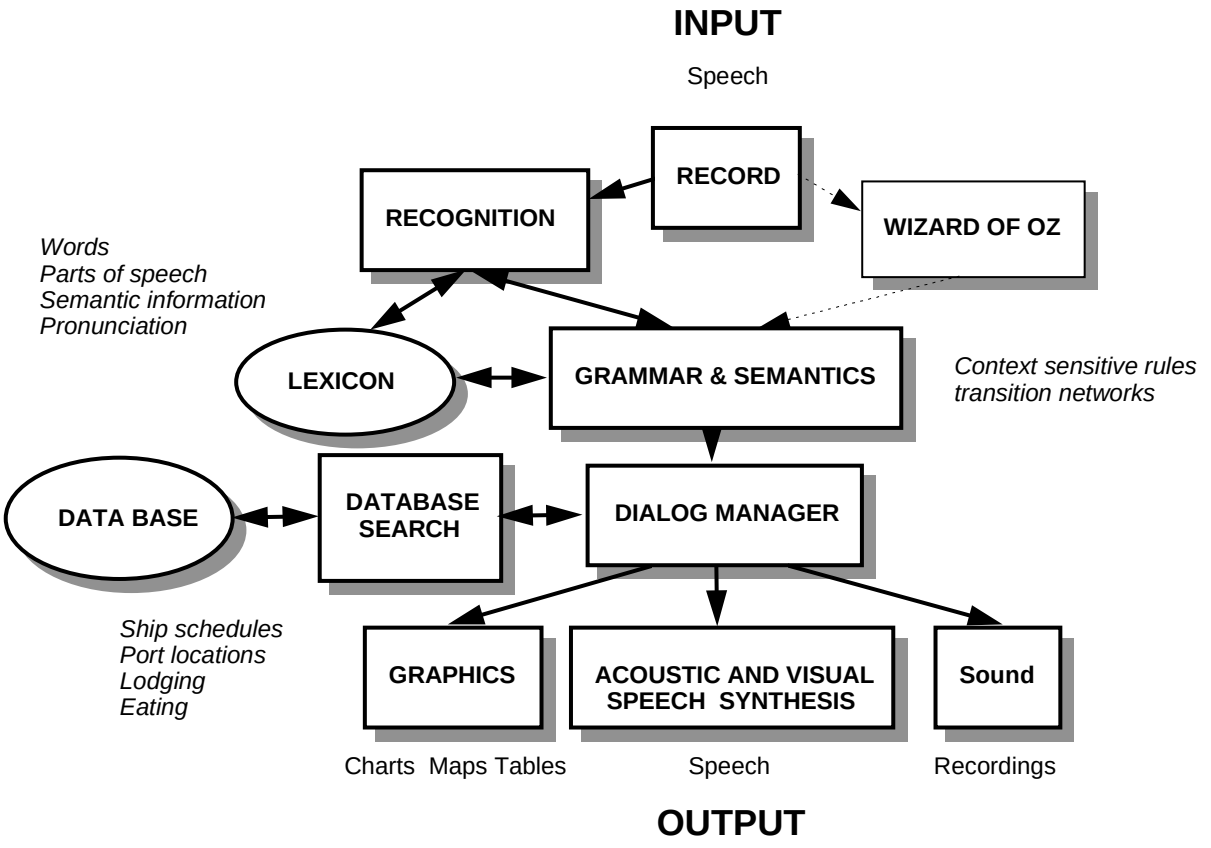

Figure 1. The modules of the Waxholm spoken dialogue system

THE NATURAL LANGUAGE COMPONENT AND DIALOGUE MANAGEMENT

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.

Some of our fundamental concepts in our natural language component are inspired by TINA, a parser developed at MIT, (Seneff, 1992). Our parser, STINA, is knowledge based and is designed as a probabilistic language model. A detailed description of the parser can be found in Carlson and Hunnicutt (1995) and Carlson et al. (1995). STINA 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 and a feature-passing mechanism to implement unification.

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 dif-

ferent 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. The topic selection based on probabilities in our system has similarities with the effort at AT&T (Gorin, 1994; Gorin et al., 1994). A different approach, also based on training, has been presented by Kuhn and De Mori (1994) in their classification approach. The work by Nowel and Moore goes one step further exploring non-word based topic spotting based on a dynamic programming technique (Nowel and Moore, 1995).

The decision about which topic path to follow in the dialogue is based on several factors such as the dialogue 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.

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. The topic prediction has been trained using a labelled set of utterances taken from the Waxholm database

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

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.

In the implementation of the parser and the dialogue management, we have stressed an interactive development environment. It is possible to study the parsing and the dialogue flow step by step when a graphic tree is built. It is even possible to use log files collected during Wizard of Oz experiments as scripts to repeat a specific dialogue, including all graphic displays and acoustic outputs.

We have added a graphical interface to the system which presents each network graphically. Both the syntax and the dialogue networks can be modelled and edited graphically with this tool. Earlier work on dialogue modelling such as the Generic Dialogue System Platform in the Danish dialogue project (Larsen and Baekgaard, 1994) has been an inspiration for this expansion

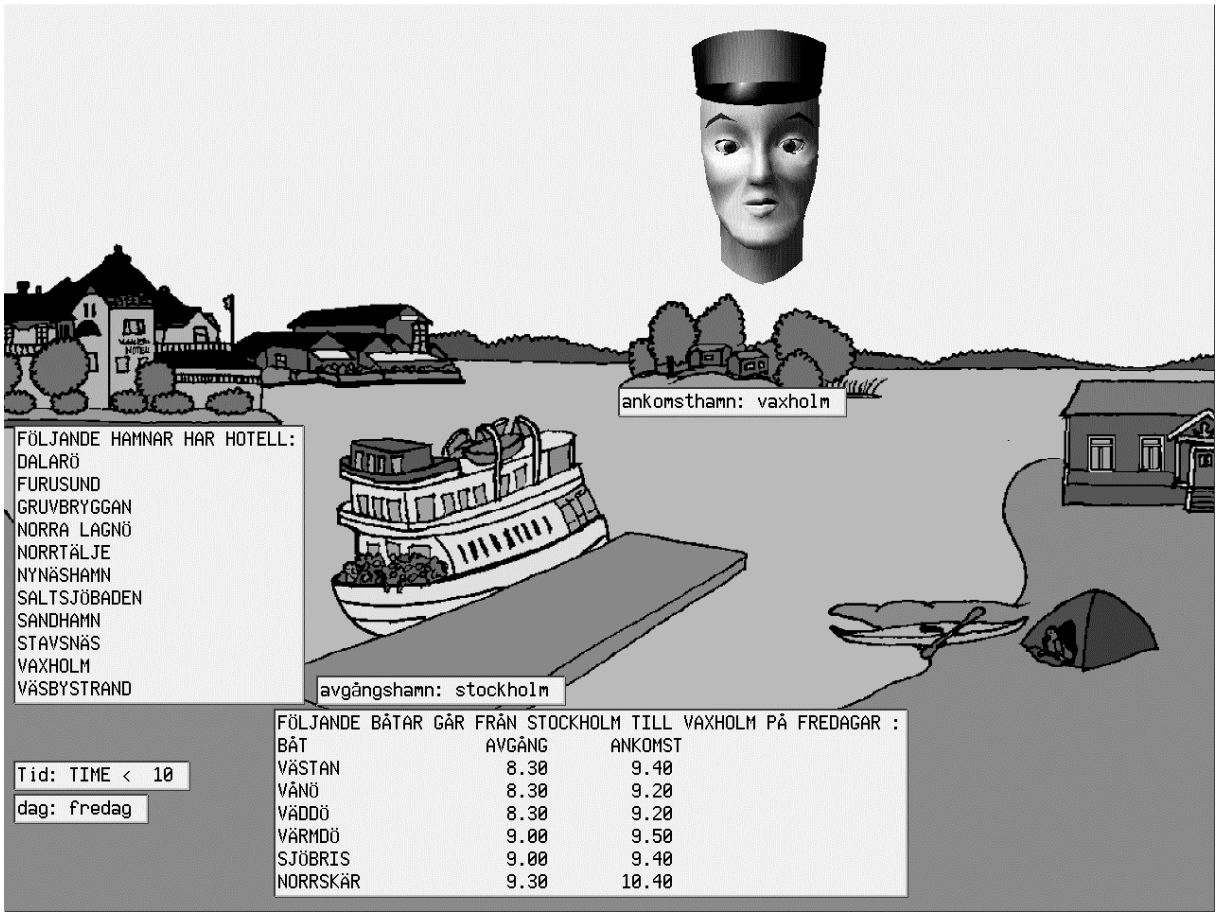

Figure 2. The graphical model of the WAXHOLM micro-world.

GRAPHICAL USER INTERFACE

The Waxholm system can be viewed as a micro-world, consisting of harbours 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.

To deal with these problems a graphical representation that visualises the Waxholm micro-world is being implemented. An example is shown in Figure 2. 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.

SPEECH SYNTHESIS

For the speech output component we have chosen our multi-lingual text-to-speech system (Carlson, Granström and Hunnicutt, 1991). The system is modified for this application. The application vocabulary has been checked for correctness, especially considering the general problem of name pronunciation (Gustafson, 1994). Furthermore we are developing multi-modal synthesis and a dialogue related prosody model.

Multi-modal synthesis

The visual channel in speech communication is of great importance, as has been demonstrated by for example McGurk (1976). A view of the face can improve intelligibility of both natural and synthetic speech, especially under degraded acoustic conditions (Le Goff et al., 1994). Moreover, visual signals can express emotion, add emphasis to the speech and support the interaction in a dialogue situation through e.g. turn-taking signals and back-channelling. This makes the use of a computer-synthesised face to create visual speech synthesis an important complement to traditional speech synthesis, especially in applications for hearing impaired people, in noisy environments and in speech based multi-modal user interfaces.

By extending the KTH rule-based text-to-speech synthesis system with a real-time animated 3D model of a human face, a system for audio-visual speech synthesis has been developed. The face is controlled from the same text-to-speech rule compiler that controls the auditory speech synthesiser. This provides a unified and flexible framework for development of audio-visual text-to-speech synthesis that allows the rules controlling the two modalities to be written using the same notation (Beskow, 1995).

The facial model

The face synthesis is based on a model developed by Parke (1982). The model consists of a three dimensional mesh with 800 vertices, connected together by about 800 polygons, which approximate the surface of a human face, see Figure 3. The shape of the facial surface is controlled by a set of parameters, which operates directly on the co-ordinates of the vertices. The topology of the polygon network remains constant.

There are about 50 parameters, which can be divided into two groups:

- expression parameters, that control the articulation and mimic of the face. These include jaw rotation, eyebrow shape and position, various mouth shape parameters etc.
- conformance parameters, that control static features in the face like for example nose length and jaw width.

A number of modifications have been made to the original model, in order to make it more suitable for speech synthesis. These modifications include introduction of a tongue and creation of a new set of parameters to control lip movements.

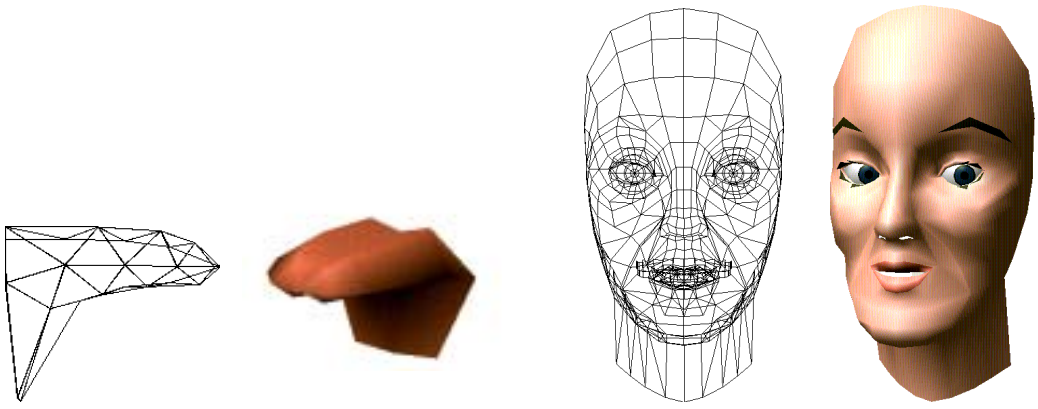

Figure 3. Wireframe and shaded representations of the tongue and the face models.

Visual prosody

The visible articulatory movements are mainly those of the lips, jaw and tongue. However, these are not the only visual information carriers in the face during speech. Much information related to phrasing, stress, intonation and emotion are expressed by for example nodding of the head, raising and shaping of the eyebrows, eye movements and blinks.

This kind of facial actions should also be taken into account in a visual speech synthesis system, not only because they may transmit important non-verbal information, but also because they make the face look alive.

These movements are more difficult to model in a general way than the articulatory movements, since they are optional, and highly dependent on the speakers personality, mood, purpose of the utterance etc.

Nevertheless, a few general rules apply to most speakers. For example, it's quite common to raise the eyebrows at the end of a question and to raise the eyebrows at a stressed syllable. There have been attempts to apply such rules to facial animation systems (Pelachaud, 1991). A few such visual prosody rules have been implemented in our visual speech synthesis system.

However, to make the face livelier, one does not necessarily have to synthesise meaningful non-verbal facial actions. By introducing random eyeblinks and very faint eye and head movements, the face looks much more alive, and becomes more pleasant to watch. This is especially important when the face is not talking, e.g. during silent periods in a dialogue application.

Recently, the visual speech synthesis module has been incorporated into the dialogue system. This is expected to raise intelligibility of the systems responses and questions. But the addition of the face into the dialogue system has many other exciting implications. Facial non-verbal signals can be used to support turn taking in the dialogue, and to direct the users attention in certain ways, e.g. by letting the head turn towards time tables, charts etc. that appear on the screen during dialogue.

The dialogue system also provides an ideal framework for experiments with non-verbal communication and facial actions at prosodic level, as discussed above, since the system has a much better knowledge of the situation than is the case in plain text-to-speech synthesis.

The dialogue prososdy model

An adequate prosody model for a spoken dialogue system must be able to explain and generate the different configurations of prosodic parameters that we observe in spoken dialogue analyses. A model and theory of spoken dialogue must be able to relate such prosodic patterns to features of the dialogue situation. Since our current prosody model is based on monologue situations with a one-sentence context window, it is our goal to develop it for use in multi-sentence-window applications and specifically to be relevant in dialogue situations.

The enhanced model, still under development, uses the same basic building blocks as our standard model (Bruce and Granström, 1993). The main extension lies in the fact that a number of gradational elements are being added, which relate to the phonetic realisation of the elements of the standard model, which are basically phonological and discrete in nature (Bruce et al., 1995).

The scope of the model is the major phrase unit, which may consist of two or more minor phrases. The division of the major phrase into minor phrases is manifested in the model by either boundary signals at the junction between them or by signs of cohesion within one or more of them.

Each phrase can be divided into a number of domains: initial juncture domain, (optionally) prefocal domain, focal domain, (optionally) postfocal domain, terminal juncture domain. As a variant, a non-focal main domain may take the place of a prefocal + focal domain.

Within each domain, each major event, i.e. in practice each turning point, is independently variable with respect to both the value and the timing of F0.

The components of the enhanced model fall into two groups,

A discrete elements

B gradational elements

The components of the standard model belong, for the most part, to group A.

The discrete elements of group A are mainly phonological/linguistic or pertain to grammatical, semantic and textual structure. These elements are typically binary. The parametric values of the gradational elements of group B mainly reflect paralinguistic factors, such as the emotional state or the regional background of the speaker, his or her attitude in the speaking situation, or other pragmatically determined factors, related for instance to the dialogue situation. The structure of the model is shown in Table 1. Each of the elements of the model can be parametrically varied, independently within each of the domains referred to earlier.

The durational parameters affect the rhythmic structure both within and across metric feet. The timing parameters affect the turning points of the basic F0 events, something which is crucial to achieving authentic sounding synthesis, and which it is hypothesised offers important cues in the perception of the nuances of natural speech. Pauses can be specified with respect to whether they are silent or filled, and whether they affect the durational relationships of the previous segments, and if so, to what extent.

There exists a close relationship between the elements in the two groups, in that those in group A are manifested and specified in terms of those in group B, that is, it is typical of the elements of B that they pertain to the *realisation* of those of A.

Although the scope of the model is restricted to the major phrase, it is part of our wider work to examine the parametrical realisations within the major phrase as seen in the wider context of a dialogue situation. Here, too, the phonetic realisation of initial or terminal juncture or signs of cohesion or lack of cohesion across a major phrase boundary may function as means to signal the structure of a succession of major phrases in a dialogue setting.

Table 1 Components of the enhanced model

A	**discrete elements**	
	tonal structure:	
	accented	accent I (HL*)
		accent II (H*L)
	focused	accent I ([H]L*H)
		accent II (H*LH)
		compound (H*L...L*H)
	juncture	initial (%L; %H)
		terminal (L%; LH%)
	grouping:	
	boundary	minor |
		major ||
B	**Gradational elements**	
	F0 phenomena	
	F0 range	
	F0 register	
	general direction of F0 movement (slope)	
	timing of F0 events	
	Duration	
	Voice source characteristics	
	Reduction phenomena	

Linking the prosody model to the parameters of dialogue structure

Having established, on the one hand, the prosodic parameters - both discrete and gradational - and, on the other, the relevant categories of dialogue structure and human dialogue management, the next task is to establish the link between the two.

To do this we have implemented the enhanced prosody model in an experimental speech synthesis system. In this we have defined special prosodic markers which can be inserted in the text. These markers represent particular settings of the gradational prosodic parameters of the enhanced model so that it is possible to select virtually any point in the multi-dimensional space created by these parameters. They can represent complex combinations like "reduced F0 range at a raised level prefocally, with extended F0 range in focus, followed by a rising terminal juncture". The combinations chosen generate patterns that we observe in our speech database and which we hypothesise are relevant indicators of dialogue-related prosodic behaviour.

By coupling the enhanced synthesis to the Waxholm dialogue system we can study the prosody of both the automatically generated speech and that of the human participants in this new environment. One hypothesis, which can then be tested, is that the limited range of prosodic variation exhibited by the human speakers in our man-machine material is correlated to the limitations in prosodic variation and expressiveness on the part of the machine dialogue partner.

Looking further into the future, the implementation of this model in an automatic dialogue system will facilitate the production of appropriate prosody once it has the means to make contextually based choices among a repertoire of prosodic possibilities.

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 (Elenius and Tråvén, 1993; Elenius and Blomberg, 1992). 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 (Ström, 1994).

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 (Blomberg, 1994). 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 merged 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.

DATA COLLECTION

Speech and text data have been collected running the system with a Wizard of Oz replacing the speech recognition module Bertenstam et al., 1995b). The subjects are seated in an anechoic room in front of a display similar to Figure 2. 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 (Blomberg and Carlson, 1993), 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 labelled 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 labelled. This kind of sound occurs when a consonant constriction is released. Vowel insertion and other extralinguistic sounds seem to be speaker specific features.

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.

ACKNOWLEDGEMENT

We thank all the people in the Waxholm project group who have been engaged in system design, data collection and data analysis. 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: 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. We also acknowledge the contribution from a related project *Prosodic Segmentation and Structuring of Dialogue*, in co-operation with Gösta Bruce and coworkes at Phonetics at Lund University. Both projects are supported by grants from The Swedish National Language Technology Program.

REFERENCES

Aust H, Oerder M, Seide F and Steinbiss V (1994). Experience with the Philips Automatic Train Timetable Information System. In: Proc IEEE workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA94), pp. 67-72.

Bertenstam J, Blomberg M, Carlson R, Elenius K, Granström B, Gustafson J, Hunnicutt S, Högberg J, Lindell R, Neovius L, Serpa-Leitao A de, Nord L and Ström N (1995a). The Waxholm system - a progress report. Proc. Spoken Dialog Systems, Vigsø, pp. 81-84.

Bertenstam J, Blomberg M, Carlson R, Elenius K, Granström B, Gustafson J, Hunnicutt S, Högberg J, Lindell R, Neovius L, Nord L, Serpa-Leitao A de and Ström N (1995b). Spoken dialog data collected in the Waxholm project. STL-QPSR, KTH, 1: 49-74. Dept. of Speech Comm. and Music Acoustics, KTH, Stockholm, pp. 49-74.

Beskow J (1995). Rule-based visual speech synthesis. In: Proc Eurospeech '95, 4rd European Conference on Speech Communication and Technology, Madrid, pp. 299-302.

Blomberg M, Carlson R, Elenius K, Granström B, Gustafson J, Hunnicutt S, Lindell R and Neovius L (1993). An experimental dialog system: Waxholm. In: Proc Eurospeech '93, 3rd European Conference on Speech Communication and Technology, Berlin, pp. 1867-1870.

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

Blomberg M and Carlson R (1993), Labelling of speech given its text representation, In: Proc Eurospeech '93, 3rd European Conference on Speech Communication and Technology, Berlin, pp. 1775-1778.

Bruce G and Granström B (1993), Prosodic modelling in Swedish Speech synthesis, Speech Communication 13, pp. 63-73.

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

Carlson R (1994). Recent developments in the experimental “Waxholm” dialog system. In: ARPA Human Language Technology Workshop, Princetown, New Jersey, pp. 207-212.

Carlson R and Hunnicutt S (1995). The natural language component - STINA. In: STL-QPSR, KTH, 1: 29-48.

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

Carlson R, Hunnicutt S and Gustafson J (1995). Dialog mangement in the Waxholm system. In: Proc. ESCA/ETRW on Spoken Dialog Systems, Vigsø, Denmark, pp. 137-140.

Clementino D and Fissore L (1993). A man-machine dialog system for access to train timetable information. In: Proc Eurospeech '93, 3rd European Conference on Speech Communication and Technology, Berlin, pp. 1863-1866.

Dalsgaard P and Baekgaard A (1994). Spoken Language Dialog Systems. In: Proc in Artificial Intelligence, Infix. Presented at the CRIM/FORWISS workshop on 'Progress and Prospects of Speech Research and Technology, Munich.

Elenius K and Blomberg M (1992). Experiments with artificial neural networks for phoneme and word recognition. In: Proc ICSLP International Conference on Spoken Language Processing, Alberta, Canada, pp. 1279-1282.

Elenius K and Tråvén H (1993). Multi-layer perceptrons and probabilistic neural networks for phoneme recognition. In: Proc Eurospeech '93, 3rd European Conference on Speech Communication and Technology, Berlin, pp. 1237-1240.

Gerbino E and Danieli M (1993). Managing dialog in a continuous speech understanding system. In: Proc Eurospeech '93, 3rd European Conference on Speech Communication and Technology, Berlin, pp. 1661-1664.

Glass J, Flammia G, Goodine D, Phillips M, Polifroni J, Sakai S, Seneff S and Zue V. Multilingual spoken-language understanding in the MIT Voyager System. In: Speech Communication. Vol. 17:1-2. pp. 1-18.

Gorin A (1994). Semantic associations, acoustic metrics and adaptive language aquisition. In: Proc ICSLP, International Conference on Spoken Language Processing, Yokohama, pp. 79-82.

Gorin A, Hanek H, Rose R and Miller L (1994). Automatic call routing in a telecommunications network. In: Proc IEEE workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA94), pp. 137-140.

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

Kuhn R and De Mori R (1994). Recent results in automatic learning rules for semantic interpretation. In: Proc ICSLP, International Conference on Spoken Language Processing, Yokohama, pp. 75-78.

Larsen L and Baekgaard A (1994). Rapid prototyping of a dialog system using a generic dialog development platform. In: Proc ICSLP International Conference on Spoken Language Processing, Yokohama, pp. 919-922.

Le Goff B, Guiard-Marigny T, Cohen M and Benoît C (1994). Real-time analysis-synthesis and intelligibility of talking faces, Proceedings of the second ESCA/IEEE Workshop on Speech Synthesis, New Paltz, New York, USA.

McGurk H and MacDonald J (1976). Hearing lips and seeing voices, *Nature*, 264, pp 746-748.

Nowel P and Moore R (1995). The application of dynamic programming techniques to non-word based topic spotting. In: Proc Eurospeech '95, 4rd European Conference on Speech Communication and Technology, Madrid, pp. 1355-1358.

Oerder M and Aust H (1994). A realtime prototype of an automatic inquiry system. In: Proc ICSLP International Conference on Spoken Language Processing, Yokohama, pp. 703-706.

Parke F (1982). Parametrized models for facial animation, *IEEE Computer Graphics*, 2(9), pp 61-68.

Peckham J (1993). A new generation of spoken dialog systems: results and lessons from the SUNDIAL project. In: Proc Eurospeech '93, 3rd European Conference on Speech Communication and Technology, Berlin, pp. 33-40.

Pelachaud C (1991). Communication and Coarticulation in Facial Animation, *Ph.D. dissertion*, University of Pennsylvania.

Seneff S (1992). TINA: A natural language system for spoken language applications. Computational Linguistics, 18(1) pp. 61-66.

Shirai K and Furui S (1995). Spoken Dialog. In: (Shirai and Furui, eds.) Special issue of Speech Communication, 15(3-4).

Ström N (1994). Optimising the lexical representation to improve A* lexical search, . In: STL-QPSR, KTH, 2-3: 113-124.