HANDS-ON SPEECH TECHNOLOGY ON THE WEB

Rolf Carlson, Björn Granström, Joakim Gustafson, Erland Levin and Kåre Sjölander
(Names in alphabetical order)

Centre for Speech Technology (CTT)
Department of Speech, Music and Hearing, KTH
SE-10044 Stockholm, Sweden
Tel.:+46 8 790 7879 Fax: +46 8 790 7854 E-mail: { rolf | bjorn | joakim_g | erl | kare }@speech.kth.se

ABSTRACT

The speech group at KTH has developed a number of speech technology tools for use in education of undergraduate students or researchers in the speech field. Many of these tools have been limited to a certain computer environment and the need for teacher guidance. The past years we have started developing a toolkit for spoken language technology that can be used over the Internet. In this presentation we want to focus on the possibilities to increase awareness of speech technology through both demonstrations and interactive experimental education, available on Internet. Examples of some of these applications will be demonstrated at the workshop.

USING THE INTERNET FOR SPEECH TECHNOLOGY DEMONSTRATIONS

Over the past three years the web has come to play an increasingly important role in our applications and demonstrations. The web is an important medium for publishing, as an application platform and as a knowledge base. In this presentation we will describe and present a number of examples where we have used the web for interactive speech technology demonstrations. One of the first efforts at CTT to use this novel vehicle for co-operative research was the creation of a web interface to our multi-lingual speech synthesiser [1]. This has made it possible to easily demonstrate the technology using any web browser anywhere. As an example, in 1996 we set up a demonstration of MIT´s dialogue system GALAXY with our web-based speech synthesis as an integrated module. Several other speech R&D sites have set up similar facilities. We have collected some examples of speech synthesis presented on the web [2].

Software for web based speech analysis has also been developed [3]. The processing is done using speech analysis tools embedded in web pages. Students can record their own speech directly from the web page, view spectrograms, make measurements etc. The expectation is that the student in an explorative and interactive process will obtain a deeper understanding of the speech signal and its variability

We have developed plug-ins for Netscape for speech recognition, where the plug-in at the client side handles the audio recording and feature extraction, while servers at CTT perform the actual recognition and optionally the natural language processing and information retrieval. The response from the server could be anything ranging from the recognised text string to information in different forms, for example web-pages, images or synthesised speech.

A modular spoken dialogue system has been developed [4], with our speech technology servers available over the Internet. The dialog system has been used in laboratory assignments in several speech technology classes in different universities in Sweden. The aim of this work has been to give students hands-on experience via a fully functioning spoken dialogue system as a teaching aid. Students were given some initial guidance on how to modify and extend the system but most of their work was unsupervised. We did an experiment running eight dialog systems simultaneously by about 20 students.

In the next sections we will describe the different applications in more detail

THE TMH SPEECH TOOLKIT

We have recently developed a toolkit to alleviate the arcane art of constructing spoken language systems. This toolkit is based on the software technology in our existing spoken dialogue system WAXHOLM [5]. We have extracted and redesigned different components such as the ones for speech recognition [6], speech synthesis [7], visual speech synthesis [8], and parsing [9] and created Tcl language [10] modules from these. This has enabled us to take advantage of the rapid prototyping and development framework, which this language fosters and to create a toolkit for spoken language technology in the spirit of the ones created at OGI (CSLUsh) [11] and MIT (Sapphire) [12]. Our toolkit has empowered us to create new applications quickly and easily based on its modules using Tcl as a glue language and also to use the accompanying Tkwidget set for graphical user interfaces. Tcl is a rather simplistic language, with many shortcomings, but we use it only for system integration and user interfaces. The modules themselves are exclusively written in C and Java. Module interfaces are string based, which makes coding, testing, and debugging simple, but these interfaces could be limiting for future applications.

The Broker Architecture

We have developed an architecture for communication between programs on different computers [13]. It consists of one central server, the Broker, which relays function calls, results and error conditions between clients and servers over the Internet. It is designed to be simple and robust.

All communication within the broker system is in text form to ensure portability and aid in debugging. There is a debugging tool to watch all these transactions. Binary data, such as speech, is sent over separate TCP connections directly from producer to consumer.

This architecture lets us keep our speech technology servers in house where they can run on powerful machines, and be updated at any time. We can also collect speech data regardless of where the applications are used.

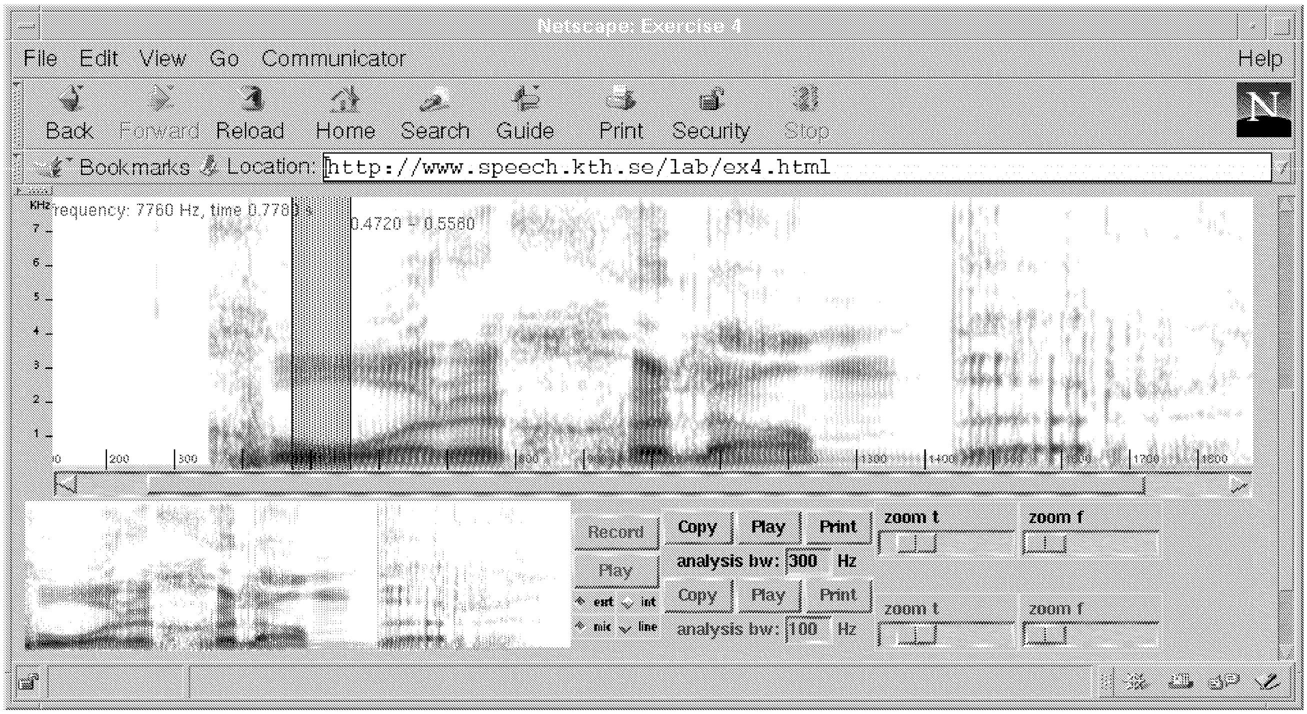
The Broker is written entirely in Java, and can therefore run on any modern computer. There are Java classes and Tcl code to support the writing of client and server modules, but the protocol is defined for any program that can communicate using the TCP protocol. We have tools that can display all communication between modules to facilitate debugging.

The Broker has been successfully used with about 50 simultaneous connections, and also between an application in Washington D.C. and our servers in Stockholm.

The full source code for the Broker system is available "as is" at http://www.speech.kth.se/proj/broker.

SPEECH ANALYSIS ON THE WEB

In our courses on speech technology we have an introductory section on basic phonetics and speech analysis. For this section we have developed a set of exercises in which students analyse their own speech in various ways. These exercises are accessed through web pages [14] in which simple speech analysis tools have been embedded as applets. In this way we have been able to supply these exercises to students both working in our lab, at Linköping University and from their home PC´s. The analysis tools were created using Tcl/Tk and an extension module developed at our lab. This makes it possible to run these tools from Netscape on three different Unix platforms and from Netscape or Internet Explorer on MS Windows. The tools can also be executed outside and independently of a web browser, but in this case they first must be downloaded and installed. The big advantages of using a web browser as a platform is that all installation issues are solved and instructions and other useful information can accompany the tools in an easily accessible way. A screen-shot of one of the exercises is shown in Figure 1. The exercises covered measurements of vowel formant frequencies, comparisons of speakers and speaking styles, Swedish word accent, and phonetic segmentation.



Location: http://www.speech.kth.se/lab/ex4.html

Frequency: 7760 Hz, time 0.778

0.4720 - 0.5560

analysis bw: 300 Hz

analysis bw: 100 Hz

Figure 1. A screen-shot of one of the speech analysis exercises running inside Netscape navigator.

SPEECH SYNTHESIS LAB

We have developed a graphical interface to our text-to-speech system [7]. This system uses a synthesis server via the broker architecture. The synthesis server that runs on our HP-UX-system is for the moment set up with five of the ten languages available in our text-to-speech system. With this set-up our speech synthesis can be used by clients anywhere in the world. The Swedish version uses a number of lexicons, including a large name lexicon constructed as part of the Onomastica project [15].

The simplest graphical interface to the synthesis server only has a text entry, a transcription field and buttons to create and play synthesised speech. If the user wants to change the prosodic pattern in the sentences, they can switch to a larger text input window where the text can be tagged with different stress levels, indicated by different colours, see Figure 4 in the section about the dialog system. Texts that have been modified in this interface can be sent to the most advanced interface where the user can edit all parameters in the formant synthesiser, see Figure 2.

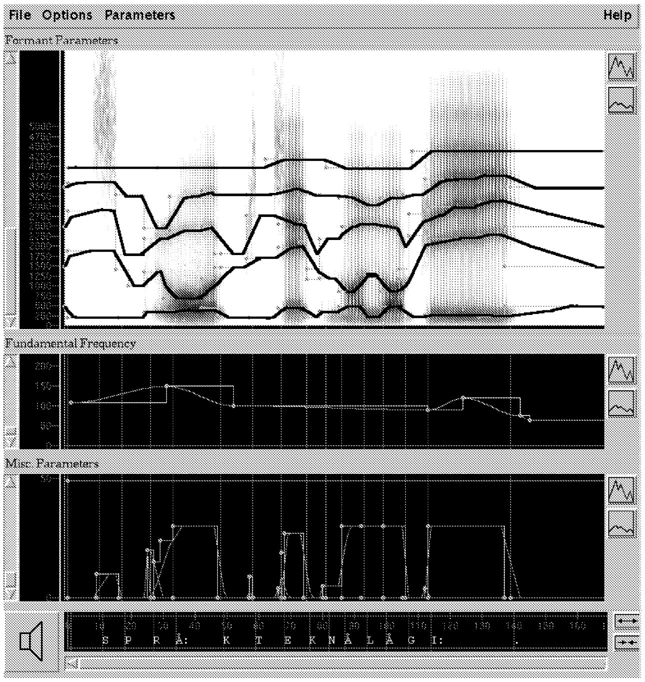


Figure 2. Combining two modules (speech synthesis generation and spectrogram) in one system.

In this interface it is possible to select which parameters to display an edit. The parameter trajectories are easily changed with the mouse and synthesis can be generated automatically whenever a control point is released. The system can also handle the parameters that are used for the multi-modal synthesis, simply by *choosing Use face for Playback* on the meny. A new window with the animated 3D-face will appear as well as new parameters that are used for the face animation. In this set-up the user can experiment with visual cues in the multi-modal speech synthesis, such as eyebrow raise at important words, or to indicate questions. It is also possible to combine the synthesiser module with speech analysis module described in the previous section, as is shown in Figure 2. The software is a new platform for speech synthesis experiments in education, and also in research**.**

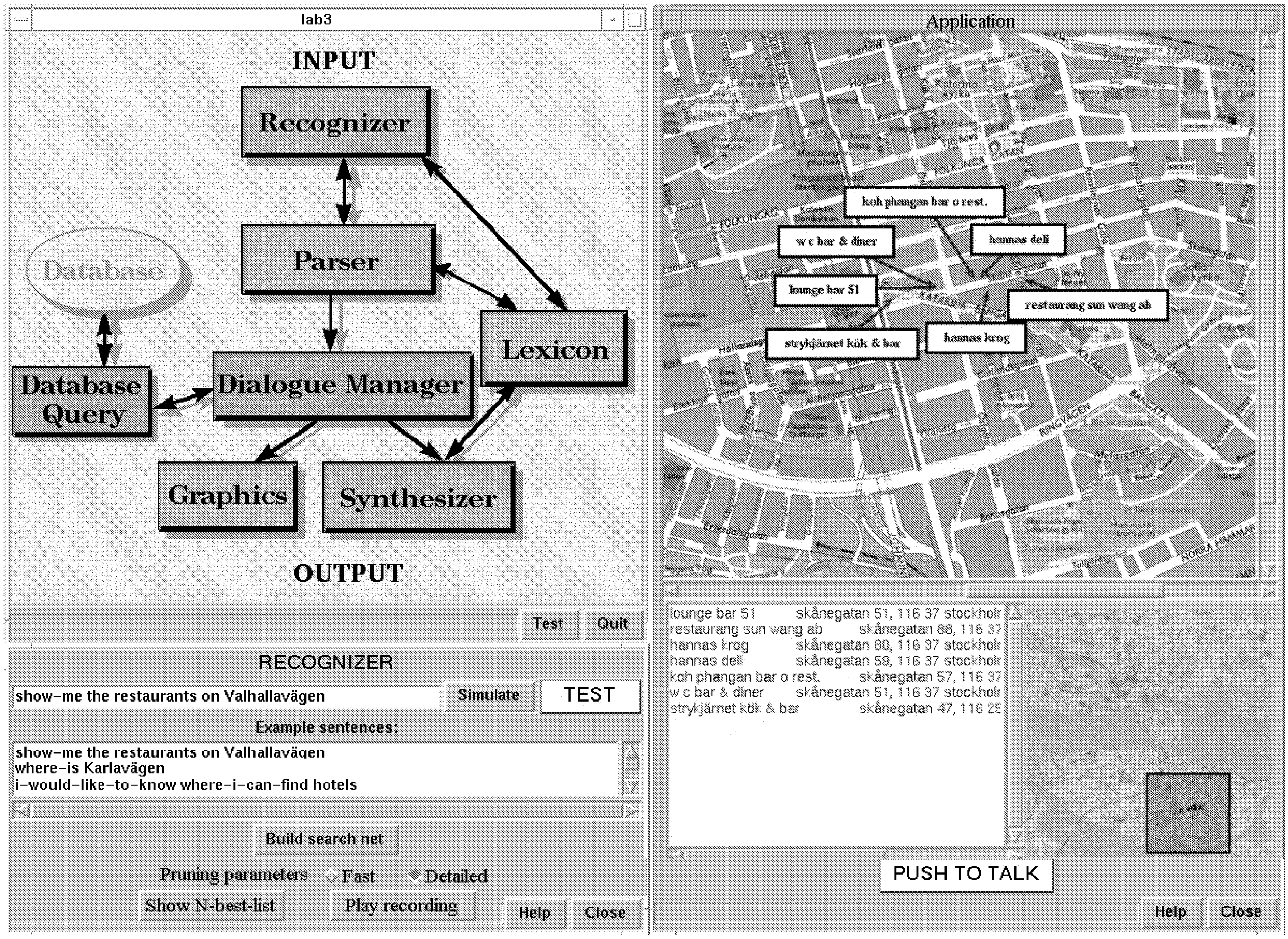
A SYSTEM FOR TEACHING SPOKEN DIALOGUE SYSTEMS TECHNOLOGY

Traditionally, spoken dialogue systems have required high expertise to design and develop. The resulting applications have been large and complicated and not always easy to modify and extend or even comprehend. Thus, teaching in the subject of spoken dialogue systems and related technologies has mostly been done in lecture format with video taped demonstrations of actual systems. Live demonstrations are typically conducted by a well-behaved PhD student, who knows which questions to ask. Students have mostly been kept at a safe distance.

The aim of this work has been to put a fully functioning spoken dialogue system into the hands of the students as an instructional aid. They can test it themselves and are able to examine the system in detail. They are shown how to extend and develop the functionality. In this way, we hope to increase their understanding of the problems and issues involved and to spur their interest for this technology and its possibilities.

The TMH speech toolkit including the broker system with distributed servers, has been used to create an integrated lab environment that can be used on Unix machines (HP-UX, Linux, Sun's Solaris, SGI's IRIX). We are planning to port it to Windows as well. The system has been used in the courses on spoken language technology given at Masters level at the Royal Institute of Technology (KTH), at Linköping University and at Uppsala University in Sweden. In this environment, students are presented with a simple spoken dialogue application for doing searches in the web-based Yellow pages on selected topics using speech, presently in the Swedish language. The system is initialised with knowledge about streets, restaurants, hotels, museums and similar services.

Results are presented using combinations of speech synthesis, an interactive map and Netscape Navigator. This application is accompanied by a development environment which enables the students to interactively study and modify the innards of the system even as it runs. Each module has its own control window, which dynamically updates to reflect the processing as it takes place. It is easy to add new information fields from the Yellow pages, assigning them semantic tags and by a simple click update the lexicon with a transcribed entry with syntactic and semantic tags. It is possible to add any words or phrases to the lexicon and then use them a few seconds later in the speech recognizer or text



show-me the restaurants on Valhallavägen

show-me the restaurants on Valhallavägen
where-is Karlavägen
i-would-like-to-know where-i-can-find hotels

```
lounge bar 51          skånegatan 51, 116 37 stockholr
restaurang sun wang ab         skånegatan 88, 116 37
hannas krog            skånegatan 80, 116 37 stockholr
hannas deli            skånegatan 59, 116 37 stockholr
koh phangan bar o rest.        skånegatan 57, 116 37
w c bar & diner        skånegatan 51, 116 37 stockholr
strykjärnet kök & bar          skånegatan 47, 116 25
```

Figure 3. A screenshot showing the control window (upper left), the dialogue application (right), and the speech recognition module (bottom left).

generator and speech synthesizer. Students can modify the results of the recognizer and parser to test how the system is affected by different inputs and errors. Thus, it is possible to control the rest of the system from each module in the chain. We have chosen to keep the modules relatively simple in this initial version.

SYSTEM MODULES

Control Window

This window shows an outline of the components of the system and how they interact. For each box the corresponding module window can be opened with a mouse click. Also, the complete dialogue application is launched from this window. There is no explicit building step involved, as all changes to the system are made incrementally. When the system runs, each box highlights when processing in the module takes place.

Speech Recognition Module

The continuous speech recognizer uses phone models trained on spontaneous speech data collected for the WAXHOLM application. As that domain (boat traffic information) differs from the current one, we use a simple class pair grammar that is based on example sentences given to the system. This set can easily be extended by the students, e.g. to incorporate new ways of formulating questions to the system. In this module, it is possible to modify and extend the grammar and also to test the recognizer stand alone, without running the complete system. Students can listen to the recording of what they said in the previous utterance and view the 10 most probable sentences suggested by the recognizer. It is also possible to use different pruning parameters in the recognition search to trade recognition accuracy for speed. Recognition output can also be edited and sent further in the system to simulate the recognizer output.

Parser Module

Parsing is either done by a statistical parser or by a simple keyword spotter. The simple grammar used by the recognizer often produces results that are hard to parse correctly, but performance using keyword spotting is still quite useful. Keywords are tagged semantically, which is used later for the database search. Results from the parser can be modified and re-sent into the system for subsequent processing.

Dialogue Manager Module

The dialogue manager module is so far a simple control loop, which activates the different modules in turn and passes information between them. Currently there are no possibilites for the students to influence the dialogue management module, but this is the main focus for work in progress, see the section on Future developments.

Database Module

The database used in the system is the publicly available web based version of the Yellow Pages [16], provided by Telia, the Swedish PTT. In the database module of the dialog system it is possible to browse all 1384 different information categories of the Yellow Pages. The search result can be edited and saved locally for faster and more secure future access. The new category will then be assigned to a semantic tag and either the name of the locally saved file or the Internet-address with the code of a certain field can be saved. The new category is added to the lexicon with automatic transcription by a simple mouse click and a new recognizer lexicon is generated by a second mouse click. The system is after this modification ready to accept questions about the new category.

Database Query Module

The module handles database queries. It translates semantic knowledge extracted by the parser module and translates this to appropriate query strings. A list with database hits is returned. It is possible to search the database by manually entering query strings.

Lexicon Module

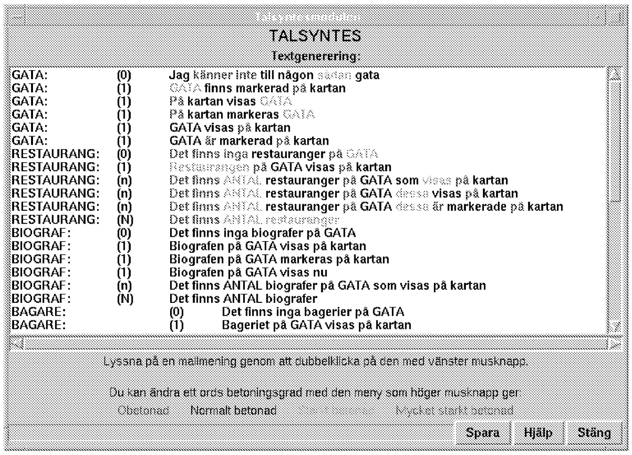
The system has a lexicon module that stores transcriptions and syntactic and semantic information for the task specific vocabulary. It can also suggest rule-based transcriptions or transcriptions from a larger general lexicon for new words entered into the lexicon. The speech synthesis module uses this task specific lexicon in the first place and, if the word is not found, uses a large general lexicon or even rules when needed. The recognizer uses only the task specific lexicon for performance reasons. Correspondingly, it is possible to check transcriptions by listening to the speech synthesis output.

Graphics Module

The graphics module displays a map with a graphical presentation of the results from the database query. Streets are highlighted and facilities are marked on the map, as shown in Figure 3.

Speech Synthesis Module

The speech synthesis module also includes response generation. The result from the database query and parser analysis is used to select a response template to fill in. There are multiple templates for each possible response type. This allows the system to choose one at random resulting in more varied system responses. It is also possible to make adjustments to the prosodic realisation of an utterance. The word stress is easily changed with the help of a pop-up meny. The word stress is marked by colouring the text, Figure 4.



```
GATA:          (0)    Jag känner inte till någon sådan gata
GATA:          (1)    GATA finns markerad på kartan
GATA:          (1)    På kartan visas GATA
GATA:          (1)    På kartan markeras GATA
GATA:          (1)    GATA visas på kartan
GATA:          (1)    GATA är markerad på kartan
RESTAURANG:    (0)    Det finns inga restauranger på GATA
RESTAURANG:    (1)    Restaurangen på GATA visas på kartan
RESTAURANG:    (n)    Det finns ANTAL restauranger på GATA som visas på kartan
RESTAURANG:    (n)    Det finns ANTAL restauranger på GATA dessa visas på kartan
RESTAURANG:    (n)    Det finns ANTAL restauranger på GATA dessa är markerade på kartan
RESTAURANG:    (N)    Det finns ANTAL restauranger
BIOGRAF:       (0)    Det finns inga biografer på GATA
BIOGRAF:       (1)    Biografen på GATA visas på kartan
BIOGRAF:       (1)    Biografen på GATA markeras på kartan
BIOGRAF:       (1)    Biografen på GATA visas nu
BIOGRAF:       (n)    Det finns ANTAL biografer på GATA som visas på kartan
BIOGRAF:       (N)    Det finns ANTAL biografer
BAGARE:               (0)    Det finns inga bagerier på GATA
BAGARE:               (1)    Bageriet på GATA visas på kartan
```

Lyssna på en mallmening genom att dubbelklicka på den med vänster musknapp.

Du kan ändra ett ords betoningsgrad med den meny som höger musknapp ger:
Obetonad   Normalt betonad   [illegible]   Mycket starkt betonad

Figure 4. The example based response generator.

Spoken Dialogue Application

The actual system that the students develop is a simple dialogue application built using the previously mentioned modules. Spoken queries are input using push-to-talk and results are presented with speech synthesis and graphics.

THE LABORATION ASSIGNMENT

The instructional environment was used in five different courses by four different departments at three universities in Sweden. These were followed by a total of 150 last-year Masters students. The students worked in groups of two and were given a list of modifications to apply to the system.

- To start with, they had to use the dialogue application in order to determine its capabilities and limitations.
- They were told to test the speech recognition module stand alone, with the explicit purpose that they should gain some insight into the limitations of current HMM based speech recognition technology. For example, regarding noise, speaking style and out of vocabulary words.
- They were told to add new fields from the Yellow Pages, new streets names from the map; and new words or phrases to the system. They had to add all new words in the lexicon with appropriate syntactic and semantic tags, and correct transcriptions.
- They had to extend the example based grammar with new constructs.
- In the text generation module, they had to insert additional response templates to handle the new facilities. They also had to experiment with different prosodic patterns in the sentences.
- Finally, the students had to demonstrate that the extended system worked according to specification.

Overall, the students were very satisfied with the system and they rated it four on a five point scale in the course evaluation. The main criticism was that they wanted to be able to make greater changes to the system and to go deeper into some of its modules. We believe that the lab environment, together with the underlying toolkit, is an important aid in giving students an understanding of spoken language technology.

FUTURE DEVELOPMENTS

Our main focus concerning the continued development of the dialogue environment is to integrate a real dialogue manager into the system. We are engaged in a joint research project together with the Natural Language Processing Laboratory, NLPLAB, at the University of Linköping, which aims at integrating the highly flexible dialogue manager [17] into the system. In this new module, a dialogue grammar based on speech act information is used together with a dialogue tree that handles focus structure.

In the current environment, the emphasis has been to give students an understanding of technology integration rather than letting them build actual new systems themselves. The latter has proven successful in a dedicated course at OGI (Colton, Cole, Novick, Sutton, 1996). Student projects, with focus on system building, will be a natural and very interesting development in our future courses.

Some work has been made to port the lab to the English language, mostly for demonstration purposes. One problem for the current system in the Yellow pages domain is the unpredictable pronunciation of Swedish street names by English speaking subjects

FUTURE USES OF THE INTERNET

We are planning to make our animated talking agent [18,19] available on the web by porting it to VRML. There are numerous possibilities for multi-modal speech in combination with other available techniques used on the web, such as translation services, language education or software agents that provide active, personal assistance [20].

We have already experimented with a very simple demonstration system that does “multi-lingual text to multi-lingual speech translation”. In this demonstration a text in English is synthesised in German, French or Spanish. It is also possible to write in any of these three latter languages and get it spoken in English. The demonstration uses AltaVista's (Systrans') web-based translation service for the translation [21]. The simple system is a 40-lines Tcl/Tk-script that uses our speech extensions and the http package in Tcl.

CONCLUSION

We believe that using the Internet will play an increasingly important role for making speech technology available anywhere for educational and co-operative purposes. Our investment in the web-based modular approach has already paid off in terms effortless portability and easy implementation of demonstrators.

ACKNOWLEDGEMENTS

This work was in part supported by the Centre for Speech Technology (CTT).

REFERENCES

[1] "Demonstrations from TMH", http://www.speech.kth.se/info/demos.html

[2] "Some examples of Speech Synthesis found on the Web", http://www.speech.kth.se/sounds/synthesis/examples.html

[3] "The SNACK Speech Visualization Module for Tcl/Tk", http://www.speech.kth.se/SNACK/

[4] K. Sjölander & J. Gustafson “An Integrated System for Teaching Spoken Dialogue Systems Technology”, Proc. Eurospeech 97, Rhodes, Greece, 1997.

[5] J. Bertenstam, M. Blomberg, R. Carlson, K. Elenius, B. Granström, J. Gustafson, S. Hunnicutt, J. Högberg, R. Lindell, L. Neovius, A. de Serpa-Leitao, L. Nord, and N. Ström, "The Waxholm system - a progress report", Proc. Spoken Dialogue Systems, Vigsoe, 1995

[6] N. Ström, "Continuous Speech Recognition in the WAXHOLM Dialogue System", STL QPSR 4/1996 pp. 67-96, Dept. of Speech, Music, and Hearing, KTH, 1996.

[7] R. Carlson, B. Granström and S. Hunnicutt, "Multilingual text-to-speech development and applications" in Ainsworth W, ed. Advances in speech, hearing and language processing. London JAI Press, 269-296, 1990.

[8] J. Beskow "Rule-based Visual Speech Synthesis" Proc. EUROSPEECH´95 Madrid, 1995.

[9] R. Carlson "The Dialog Component in the Waxholm System", Proc. ICSLP´96, Philadelphia, USA, 1996.

[10] J. K. Ousterhout, "Tcl and the Tk Toolkit." Addison Wesley, ISBN: 3-89319-793-1, 1994.

[11] S. Sutton, J. de Veilliers, J. Schalkwyk, M. Fanty, D. Novick, and R. Cole, "Technical specification of the CSLU toolkit," Tech. Report No. CSLU-013-96, Center for Spoken Language Understanding, Dept. of Computer Science and Engineering, Oregon Graduate Institue of Science and Technology, Portland, OR, 1996.

[12] L. Hetherington and M. McCandless. "SAPPHIRE: An extensible speech analysis and recognition tool based on Tcl/Tk." Proc. ICSLP '96, Philadelphia, 1996.

[13] "The Broker Architecture at TMH", http://www.speech.kth.se/proj/broker/

[14] "Speech Analysis Exercises", http://www.speech.kth.se/labs/analysis/

[15] Gustafson, J. (1994). "ONOMASTICA - Creating

a multi-lingual dictionary of European names", In FONETIK ´94, Papers from the 8th Swedish Phonetics Conference, May 24-26, Lund, Sweden, pp. 66-69.1-94

[16] "Gula sidorna (Swedish Yellow Pages)", http://www.gulasidorna.se/

[17] A. Jönsson, "A Model for Dialogue Management for Human Computer Interaction", Proc. of ISSD´96, Philadelphia, pp 69-72, 1996.

[18] "MULTIMODAL SPEECH SYNTHESIS", http://www.speech.kth.se/multimodal/

[19] Beskow, J. "Animation of Talking Agents", In Proceedings of AVSP'97, ESCA Workshop on Audio-Visual Speech Processing, Rhodes, Greece, September 1997.

[20] "Agent Group at MIT media lab", http://lcs.www.media.mit.edu/groups/agents/research.html

[21] "Alta vista's (Systrans') web-based translation service"http://babelfish.altavista.digital.com/cgi-bin/translate?