Creating web-based exercises for spoken language technology

Joakim Gustafson, Kåre Sjölander, Jonas Beskow, Rolf Carlson and Björn Granström

Department of Speech, Music and Hearing, KTH
{joakim_g, kare, beskow, rolf, bjorn}@speech.kth.se

ABSTRACT

This paper describes the efforts at KTH in creating web-based exercises for speech technology. The World Wide Web was chosen as our platform in order to increase the usability and accessibility of our computer exercises. The aim was to provide dedicated educational software instead of exercises based on complex research tools. Currently, the set of exercises comprises basic speech analysis, multi-modal speech synthesis and spoken dialogue systems. Students access web pages in which the exercises have been embedded as applets. This makes it possible to use them in a classroom setting, as well as from the students' home computers.

1 INTRODUCTION

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. Earlier, such tools were often developed for a certain computer environment and needed teacher guidance because of their complexity. When the number of speech technology students started to grow, during the last couple of years, the need for more flexible solutions became apparent. One important aim has been to provide dedicated instructional software instead of research tools. Recently, we have added issues like platform independence and intuitive user interfaces to our design goals. Another one is the possibility to use our software over the Internet. The aim is to free the students from the need of using a particular computer at a particular time and place.

Our courses on speech technology include an introductory section on basic phonetics and speech analysis. A set of exercises for this section has been developed in which students analyse their own speech in various ways. An interactive tool for working with parametric speech synthesis has been developed. The tool facilitates editing of parameter tracks, and it provides real-time feedback of the synthesised speech. It serves as an interface to KTH's multilingual rule based synthesis system, and can be used to control a formant synthesiser as well as a 3-D "talking head". The integrated lab environment GULAN [1] for spoken dialog systems has seen further development in the area of dialogue management [2] in co-operation with the NLP lab at the University of Linköping. The system has also been redesigned for web deployment.

2 IMPLEMENTATION

A speech technology toolkit that serves as a basis in the creation of spoken language systems has been developed. A number of programming languages are used in the modules of our toolkit. C/C++ is used where maximum performance is required, for example, for the speech recognition and speech synthesis engines. The Tcl/Tk language is used for user interfaces and as a glue language in some modules. It was chosen because it is platform independent and makes it simple to integrate existing modules and applications as well as changing and extending these. It also has powerful and easy to use networking facilities. The main drawbacks of the Tcl language are its execution speed and its primitive syntax. However, this can be overcome by implementing complex and time critical code in a more powerful language. Also, it is possible to run scripts which are embedded in web pages and which download quickly because of their relatively compact text representation. In all, this is an ideal solution for computer based instruction and distance learning.

A new addition is the design of an architecture for communication between programs on different computers using the Internet, the Broker Architecture [3]. The Broker Architecture relays function calls, results and error conditions between modules in text form over standard TCP internet connections. The Broker is written in Java and the modules themselves are written in a number of languages.

3 SPEECH ANALYSIS MODULE

One of our extensions to the Tcl language is the Snack speech analysis module [4]. It provides a uniform interface to the audio hardware on a number of platforms, adding commands to play, record and manipulate sound in many audio formats, as well as disk I/O in common audio file formats. Also, it has streaming audio capabilities which makes it easy to create client/server audio applications. There are commands that allow the visualisation of sounds using waveforms, spectrograms, and spectrum sections. The Snack module serves as a basis when creating customised recording tools, speech analysis applications, audio annotation tools, demonstrators, and the like. The module has a powerful and intuitive way of handling sound as objects. A spectrogram object connected to a sound object will update automatically and in real time as the sound data

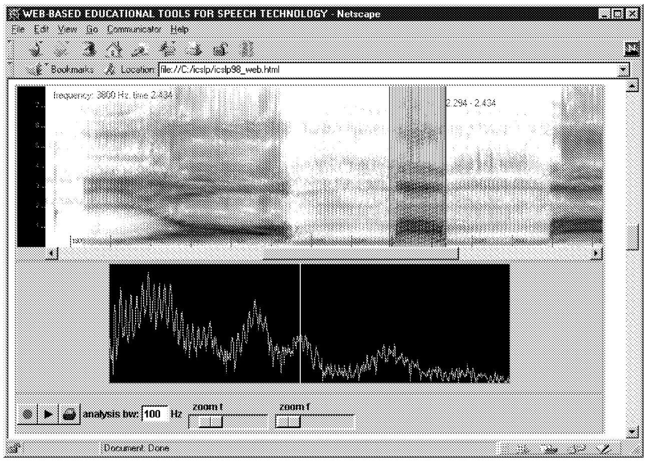

WEB-BASED EDUCATIONAL TOOLS FOR SPEECH TECHNOLOGY - Netscape
File Edit View Go Communicator Help
Bookmarks
Location: file://C:/icslp/icslp98_web.html
frequency: 3800 Hz, time 2.434
2.294 - 2.434
analysis bw: 100 Hz
zoom t
zoom f
Document: Done

Figure 1. A screen-shot of one of the speech analysis exercises.

changes. The modules also support postscript printing in order to create hard copies or, for example, to make it possible to create illustrations. Currently, it is possible to write platform independent scripts which run on Unix (Linux, Solaris, HP-UX, IRIX) and Windows95/NT using the Snack module. It is also possible to run scripts embedded in web pages through the use of the Tcl plug-in. The Snack module, in source and binary format, can be freely downloaded from http://www.speech.kth.se/snack/ An example of how this module can be used follows:

```
#!/usr/local/bin/wish
package require snack
sound snd
pack [spectrogram .s -sound snd -height 200]
pack [button .r -text Record -com {snd record}]
pack [button .t -text Stop -command {snd stop}]
```

The example creates a simple real time spectrogram application. A sound object called **snd** is created, which is empty initially. Next, a spectrogram is created, that is linked to the sound. And finally two buttons labeled Record and Stop. When clicked, these buttons will execute the commands **record** and **stop,** respectively, of the sound object. As the recording commences the spectrogram will update in real time to reflect the changing contents of the sound object. There are numerous options to handle analysis bandwidth, scales, and similar properties. The example script could easily be extended with for example the ability to play a recording:

```
pack [button .p -text Play -command {snd play}]
```

In order to be able to save the recording the following line could be added:

```
pack [button .w -text Save -com {snd write
[tk_getSaveFile]}]
```

Also, the script above would run without modification if embedded in a web page, except for the save function that would need special privileges.

3.1 Speech Analysis Exercises

In our courses on speech technology we have an introductory section on basic phonetics and speech analysis. For this section a set of exercises were developed in which students analyse their own speech in various ways. These exercises are accessed through web pages, in which simple speech analysis tools have been embedded as small applications (applets) dedicated to the task at hand (http://www.speech.kth.se/labs/analysis/). In this way it was possible to make these exercises available to students working in our laboratory, at Linköping University, as well as those working from their home computers. The big advantage of using a web browser as a platform is that all installation issues are solved, except for the download and installation of plug-ins. Instructions and other useful information can also accompany the tools in a natural and easily accessible way, using HTML. 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.

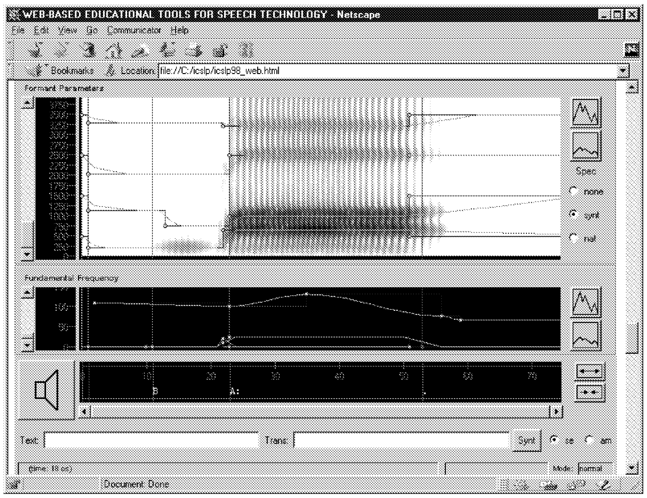
3.2 Using Snack in general applications

The object oriented approach makes Snack easy to use when building spoken dialogue systems, where a sound object can be associated to one or several speech recognition engines [5]. As soon as the sound is changed the associated recognisers automatically generate n-best-lists of utterances that are sent to a pre-defined function.

Another key feature is the possibility to use the Snack package in conjunction with software developed at other sites, such as the CSLU Toolkit [6]. Snack is currently in use at a number of sites, some of which have also contributed in the development process or extended it for internal use. A prime example of a tool using Snack is Transcriber [7] from DCE/CTA and LDC, for annotation of broadcast material.

4 SPEECH SYNTHESIS TOOL

An interactive tool for working with parametric speech synthesis has been developed. This tool facilitates editing of parameter tracks, and it provides real-time feedback of the synthesised speech. It serves as an interface to KTH's multilingual rule based synthesis system [8], and can be used to control a formant synthesiser as well as a 3-D "talking head" [9]. The parameter tracks, generated with KTH's rule based system, are also being used to control a Mbrola diphone synthesiser with a Swedish voice together with an animated agent [5]. The synthesis tool can run either as a stand-alone application or as an applet in a Web browser. It has been used in research and education. It gives full control over all parameters involved in the formant synthesis process, including formant frequencies and bandwidths, fundamental frequency and voice source parameters. The user can select a language (Swedish, French, American English or German) and synthesise arbitrary text, either in orthographic or phonetic mode. Once the phonetic transcription has been generated, the



Location: file://C:/icslp/icslp98_web.html

Spec: ○ none ● synt ○ nat

Text: Trans: ● se ○ am

(time: 18 cs)

Mode: normal

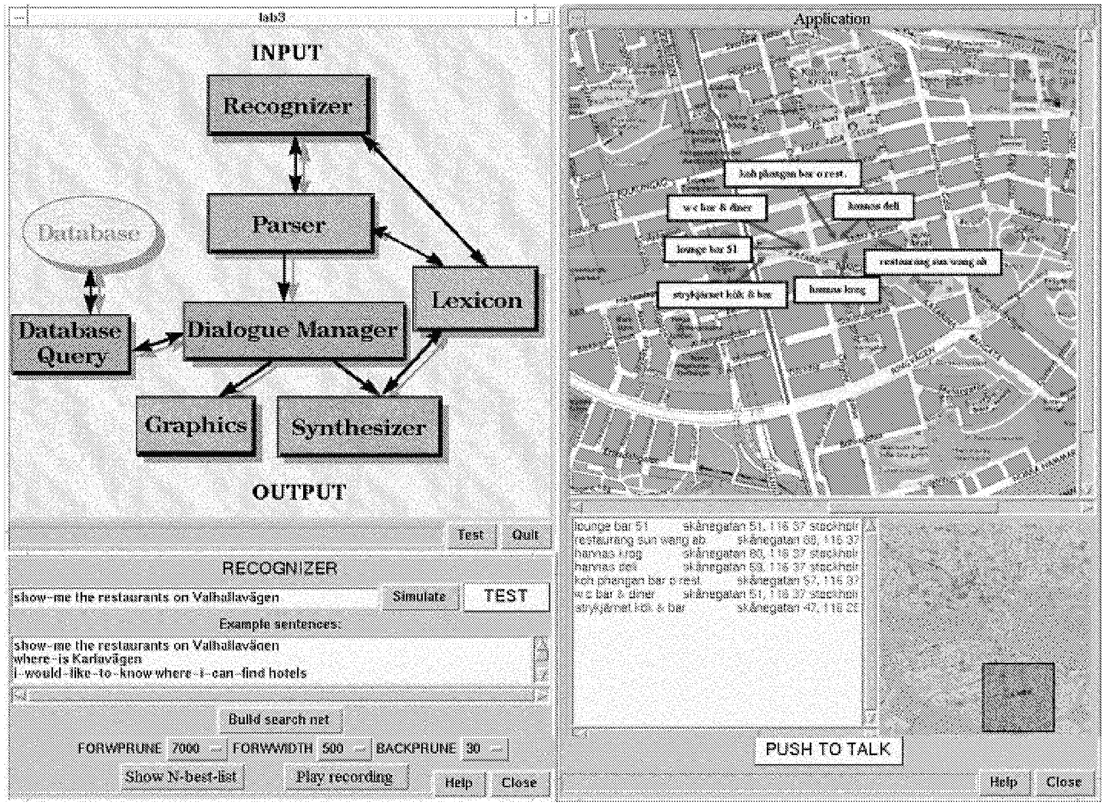
Figure 2. A screen-shot of the speech synthesis tool Veiron.

synthesiser produces the control parameter tracks in a two-step process: First, the phonetic rules generate a series of control points for each parameter that define a target track. Next, filters are applied to the target track to create a smoothed continuous track to be output to the synthesiser. The filter type and coefficients may differ between parameters, and the filter coefficients for a given parameter may vary in time, under the control of another parameter.

The main interface, shown in Figure 2, consists of a number of panels that display the parameter tracks, a time scale, a menu bar, a horizontal scrollbar and a status bar. Each of the panels has an associated value-scale and controls for vertical zooming and scrolling. The panels are stacked and aligned vertically, in such a way that all panels share the same time scale, and horizontal scrolling affects all panels. Typically, related parameters or parameters of the same unit are displayed together in one panel. The default configuration contains three panels, displaying formant parameters, fundamental frequency and source parameters respectively. The parameter tracks in the formant panel can be overlaid on top of a spectrogram. Parameter tracks are edited in an intuitive way by dragging the control points. Control points can be freely inserted or deleted, and segment durations can be lengthened or shortened using a time scale at the bottom of the display.

4.1 Speech Synthesis Exercises

Students are given a number of tasks to accomplish using the editing tool. The first task is to change the identity of the consonant in a synthesised CV syllable by manipulating the formant transitions. For example, to change /ba/ to /da/, the transitions of the second and third formant into the vowel part will have to be changed from rising to falling. In the second task, the students apply the knowledge gained in the previous speech analysis exercises, where they study the pitch contour of Swedish tones. Using a set of minimal pairs with respect to tone, the task is to synthesise the first word in the pair and manipulate the F0 contour to arrive at the other word. A similar exercise involves changing the meaning of a word by modifying vowel length, vowel quality and stress. The last task is to experiment with prosodic modifications at sentence level, such as changing a statement into question.



show-me the restaurants on Valhallavägen

Example sentences:

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

FORWPRUNE 7000 FORWWIDTH 500 BACKPRUNE 30

lounge bar 51	skånegatan 51, 116 37 stockholr
restaurang sun wang ab	skånegatan 69, 116 37
hermans krog	skånegatan 60, 116 37 stockholr
hermans 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 2[illegible]

Figure 3. A screen-shot of the educational dialogue system Gulan.

Formant based synthesis is sometimes compared to more commercially popular synthesis methods based on concatenation of segments recorded from natural speech. Concatenation based synthesis can offer high voice quality, but is limited in flexibility. Typically, only pitch and duration can be altered freely. In contrast, we feel that formant based synthesis has a significant pedagogical value. By using a parametric synthesis paradigm based on a familiar phonetic representation controlled from intuitive graphical interface, exercises can be designed that provide the students with a deeper understanding not only of fundamental speech synthesis techniques, but also about acoustic-phonetic correlates in general.

An important feature of the parameter based speech synthesis is the possibility to add new parameters for the animated talking head, so that these can be edited in the same framework as the audio parameters. This feature will be used in a dialogue lab assignment at MiLaSS, the European summer school on Multimodality in Language and Systems. In this assignment the students will be divided into groups, generating one half of a computer-computer dialogue. They will use our tools to generate synthetic speech with appropriate prosody, facial gestures and head movements.

5 EDUCATIONAL DIALOGUE SYSTEM

The educational dialogue system GULAN [2] has been redesigned into an application that is accessed through a web page. In this dialogue system users can make simple queries in the web-based Yellow Pages on selected topics using speech. Results are presented using a combination of synthesised speech and an interactive map. Our aim is to give the students hands-on experience by letting them use the system on their own, examining it in detail, and extending its functionality. In this way, we hope to give them an understanding of the problems and issues involved in building dialogue systems and to spur their interest in spoken dialogue technology and its possibilities. Recently, the system has been redesigned with an improved dialogue manager described in [3]. The system can also make use of the broker for modules such as recognition and synthesis if it must run on slow computer.

5.1 Dialogue System Exercises

The students were given a set of tasks to complete. First of all they had to use the system in order to figure out what it could and could not do. This also included experimenting with the speech recognition component itself in order to understand its current limitations regarding, for example, speaking style, vocabulary, and grammar. They could also change the pruning parameters used in the recogniser and re-generate n-best-lists of probable utterances for the last user input, thus making it possible to study their effect on speed, quality and number of hypothesis generated. The principal task was to extend the system with new fields from the Yellow Pages. New words and phrases had to be added to the lexicon and the grammar had to be modified accordingly. This is an interactive process where the students can listen to the transcriptions using the text-to-speech system. Immediately after they have loaded the updated lexicon into the running recogniser they can use the new words. They also have to extend the text generation capabilities in order to handle the new fields. The students could also modify the prosodic patterns of the synthesised responses.

6 CONCLUSIONS AND FUTURE WORK

In this paper, some our efforts in creating web-based exercises for speech technology have been presented. It has been useful to liberate our students from the need to use certain computers or special laboratory set-ups at certain hours. In fact, it has become necessary since the total number of students has risen from about 20 to 200 in two years and one of our courses was given at Linköping University.

Much remains to be done, but the basic framework has shown its strength. Our current systems will be continuously developed and extended. The speech analysis module will be expanded in order to make re-synthesis possible in conjunction with the speech synthesis tool. The educational dialogue system will be improved. Modules for multi-modal synthesis and prosodic analysis will be added, as well as dialogue dependent speech recognition and speech synthesis.

We believe that using the Internet will play an increasingly important role in 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 of effortless portability and easy implementation of demonstrators[5]

7 REFERENCES

1. Sjölander, K., and Gustafson, J. (1997) An Integrated System for Teaching Spoken Dialogue Systems Technology, in Proceedings of Eurospeech '97.
2. Gustafson, J., Elmberg, P., Carlson, R., Jönsson, A. (1998) An educational dialogue system with a user controllable dialogue manager, in Proc. of ICLSP´98.
3. Lewin E., (1998) The Broker Architecture, http://www.speech.kth.se/proj/broker/.
4. Sjölander, K., (1997) The Snack Sound Visualization Module, http://www.speech.kth.se/snack/.
5. Gustafson, J., Lindberg, N. and Lundeberg, M. (1999) The August spoken dialogue system, to be published in proceedings of Eurospeech'99.'
6. Sutton, S. et a. (1998) Universal Speech Tools: The CSLU Toolkit, in Proceedings of ICLSP´98.
7. Barras, C., Geoffrois, E., Wu, Z. and Liberman, M (1998) Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech, First International Conference on Language Resources and Evaluation. http://www.etca.fr/CTA/gip/Projets/Transcriber/
8. Carlson, R., Granström, B., and Hunnicutt, S. (1982) A multi-language text-to-speech module, Proceedings of ICASSP '82, Paris, Vol. 3, pp 1604-1607, 1982.
9. Beskow, J. (1995) Rule-based Visual Speech Synthesis, in Proceedings of Eurospeech '95.