

Speech Communication and Speech Technology



*Björn Granström
Professor in
Speech Communication*



*Rolf Carlson
Professor in
Speech Technology*

The speech communication and technology group is the largest within the department. The group engages about 35 researchers and research students, a few of them working part-time. The group includes CTT, the *Centre for Speech Technology*, which was established in 1996. The third phase started July 1, 2001. The organisation of CTT is presented on page 11?xxx.

Activities in the speech group, including CTT, cover a wide variety of topics, ranging from detailed theoretical development of speech production models, through phonetic analyses to practical applications of speech technology. Several theses have been presented during the year spanning a range of research topics including audio-visual speech synthesis, voice analysis, multimodal dialogue systems and speech recognition.

Spoken dialogue

A major focus of CTT is research on multimodal dialog systems. The objective is to study speech technology as part of complete systems and the interaction between the different modules that are included in such systems. These systems have been the platform for data collection, data analysis and research on multimodal human-machine interaction.

The AdApt system, a multimodal dialogue system for information on apartments for sale in Stockholm, has been evaluated during the period using the PARADISE framework. The evaluation of a conversational system includes new challenges compared to the standard methods for frame-based dialogue systems. It is not always easy to measure task success since the task description might have to be generated based on the current dialog status.

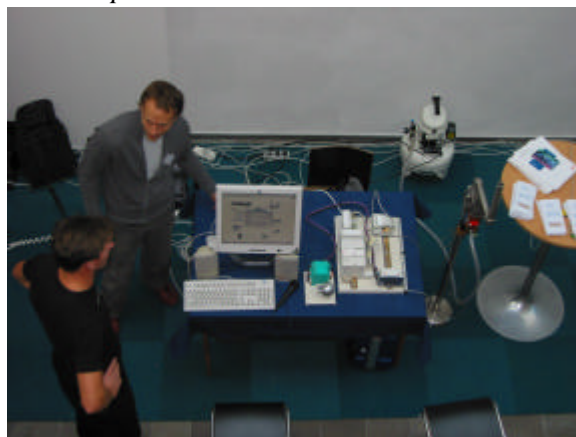
With the limitations of current speech technologies, both for recognition and understanding and for speech generation, the interest in “real” systems has led to an increased awareness of the problems raised by system errors, especially in recognizing user input, and the consequent confusion that such errors may lead to for both users and the system itself during the dialogue. The need to devise better strategies for detecting problems in human-machine dialogues and then dealing with them gracefully has become paramount for spoken dialogue systems. Several efforts have been carried out during the year along these lines. The new Higgins project will specially focus on error handling. Several data collections on error handling have been performed as a foundation for the project: a 16 subject human-human error handling study; an 8 subject human error detection study and a 24 subject study on how people describe locations in a virtual environment. The results clearly illustrate that different knowledge sources (such as confidence scores, syntactic structure and context) can be used to detect errors in recognition and react to them in an appropriate way.



The Higgins virtual environment

Mobile services and ubiquitous computing is addressed in the AlltiAllo project. This work focuses on the development of a generic adaptive system in which new services can be integrated. A first baseline system has been built in an industrial environment in which the ABB Aspect Integrator Platform is integrated with the PipeBeach Voice Web product. During 2003 a project called WiseTech (Wireless Service Technician) is running at ABB Corporate Research in which CTT is involved. Another focus in the AlltiAllo project concerns a

reception application also described in the section *Speaker characteristics* below.



An AlltiAllo experimental setup.

Linguistic processing

In addition to dialog modelling in the presented applications, research is also carried out on other general issues such as semantic modelling and also the development of lexical structures for speech technology areas. Data-driven syntactic analysis has been addressed focussing on methods and applications for Swedish. The work is now continued in the project “Boundaries and groupings - the structuring of speech in different communicative situations.” One of the goals of the project is to model the prosodic structuring, using existing as well as new speech corpora. Production and perception studies are used in parallel with automatic methods developed for analysis, modelling and prediction of prosody. The model is perceptually evaluated using synthetic speech.

Analysis of speech data from the AdApt and Higgins spoken dialogue systems has led to the development of a new robust semantic interpreter that mixes rule-based parsing with automatic techniques for managing ungrammatical and noisy input. Analysis of the data from a semantic perspective shows a need for both deep, structured semantics and flat key-value structures, something that is now built into the interpreter.

Speech recognition and databases

We see an expanding interest in studies on speaker variability, especially in the context of



Recording the SpeeCon database in the outdoor condition.

speaker independent/speaker adaptive recognition. Large text corpora are increasingly important for language technology developments. We have participated in several large efforts to build telephone speech databases such as the EU SpeechDat-project. In the EU project SpeeCon, we have collected a multi-microphone Swedish database, recorded in different environments. The database consists of material from 30 to 45 minute recording sessions by 600 speakers of which 50 are children.

We also developed several databases primarily intended for speaker verification research. The most recent, Per, contains 52 speakers recorded during several months

through different channels: fixed and mobile telephone and wideband microphone with a parallel video recording. The background speech model is based on another 79 speakers, 51 male and 28 female.

A large text corpus has been collected, containing 150 million words for use in e.g. language model experiments.

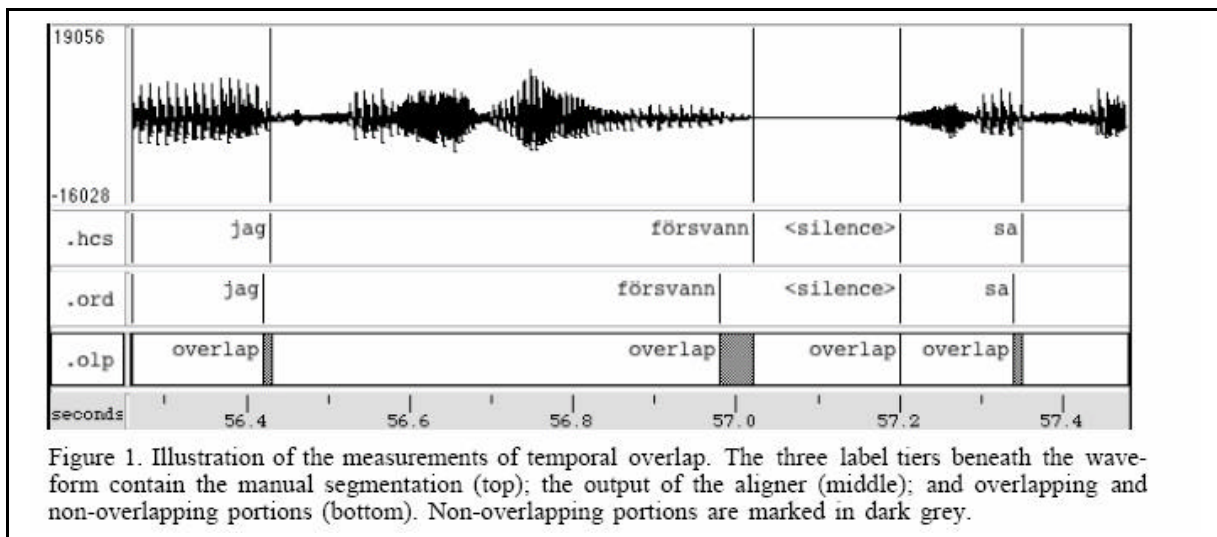
In the EU PF-Star project 200 children, 4 to 8 years old, have been recorded. This corpus is used for research on recognition of children's speech, described below in the section *Speaker characteristics*. In this project 40 Swedish children speaking English were recorded for experiments with recognition of non-native pronunciations in e.g. language training applications.

Our tool for automatic segmentation of speech has been compared to human segmentation on a (10 ms) frame by frame level. The word level segmentation precision was 90% for spontaneous speech and 95% for read speech. The human precision was 95% and 97% respectively. An example can be seen in the figure below.

A new Swedish large vocabulary speech recogniser for 30 000 words has been developed. It is based on Weighted Finite State Transducers (WFSTs). FSTs make it possible to use a unifying framework for all the different layers of the recogniser from the acoustic to phonetic layer to the language model. It works in real time with incremental continuous output.

Our fast phonetic recogniser based on Artificial Neural Networks has been further developed within the Synface project.

Regarding robust recognition a thesis work has shown that using an auditory model for the primary speech analysis works better for



phoneme recognition over the telephone than a standard recogniser (MFCC). Another thesis study on speech recorded in the fighter aircraft J39 Gripen (from 1 to 9 g) showed that adaptation to the g-level increased the speech recognition rate.

Speech production models

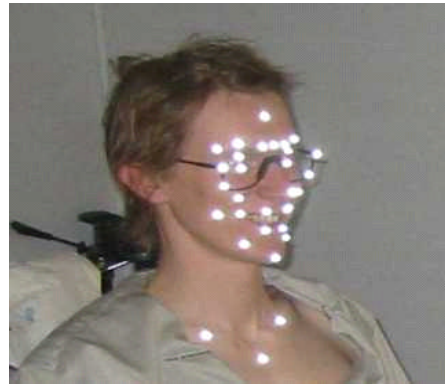
Our work on improved models of the voice source and its interaction with the vocal tract has led to a detailed understanding of the mechanisms involved. Data, in terms of the new model, on variations in natural speech have also been accumulated, both concerning linguistically motivated variations and variations among speakers.

Several ways of describing the vocal tract are being investigated, including a full 3D model. Reliable articulatory reference data still seem to be the most severe bottleneck. Both direct and indirect methods of data collection have been/are being investigated. In an effort to combine our work on the 3D-articulatory model with the talking head development we have recoded a single speaker database with combined 3D motion capture data, Qualisys and 2D mid-sagittal EMA data.

New tools for developing, controlling and playing alternative 3D models in the MPG4 standard have been developed, for easier sharing of results – primarily in the EU PF-Star project. In this project we are studying the visual aspects of expressive. We have made recordings, analysis and modelling of both traditional emotional expressions and conversational signals for e.g. feedback and turn-taking. The interaction of the non-linguistic expressions with the simultaneous phonetic articulation poses an especially interesting problem.

Speaker characteristics

The PER project is an effort to build an automated entrance receptionist (PER - Prototype Entrance Receptionist). It was operating in the central entrance at the previous location of the department and will be re-installed in the entrance to our new offices. The purpose is to create and experiment with alternative speech-based means of controlling access to the premises. Speaker verification is used for identification of employees and a dialogue system will handle the communication with occasional visitors. Text dependent and text



Recording of conversational interaction, including Qualisys motion capture.

independent algorithms are run in parallel which raises the performance of any single technique.

A large effort has been devoted to collection of an evaluation corpus. The corpus will also be used for research on channel compensation and long-term speaker variation, since the same speakers in this corpus have also been recorded in previous corpora. In the speaker verification domain, we are also engaged in the European COST 275 project.

One activity in the European PF-Star project is automatic recognition of children's speech. Children would have much use of voice for computer interaction, especially before they have acquired reading and writing skills. However, not as much research effort has been devoted to this user category as for adults. In the first year of the project, we have collected a corpus of 200 children in the age range between 4 and 8 years and started baseline recognition experiments. Preliminary results support those of previous work which exhibit significantly higher error rates for children than for adults. Methods are now being tried to improve the accuracy, including Vocal Tract Length Normalisation (VTLN) and adaptation of models trained on adult speech.

In our text-to-speech project, we have increased the efforts on different speaking styles. Both speaker variation and synthesis of attitudes, emotions and reduced speech are studied. Our long-term efforts on improved prosodic models and segmental synthesis continue.

Research on discrimination between speech and music has resulted in an HMM-based algorithm that takes account of the different segmental structure in speech and music. This method is significantly more accurate than previous methods.

Tools for education and prototyping

Our work on new tools continues. The CTT-toolbox has been used in many spoken language dialogue experiments inside and outside CTT.

The possibility of fast prototyping based on modules has proven useful in several projects.

For teaching it has resulted in a new set of student labs in speech technology. An interactive dialogue system was created in which students can change and expand the system functions. A new framework for speech synthesis is the topic of another lab.

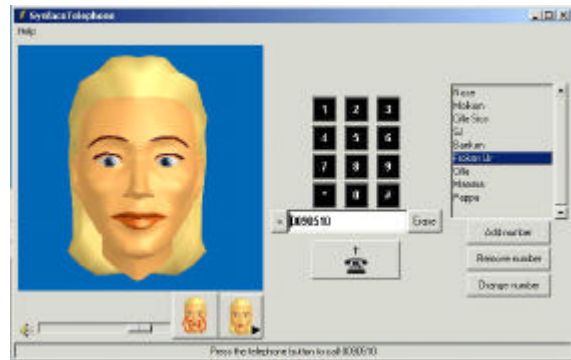
Multimodal speech synthesis

The audio-visual face synthesis project has attracted considerable attention. The synthesis is now used in many of the demonstrators under development. Strategies for articulatory synthesis are under development. The expansion of the model to the internals of the speech production apparatus is well under way and will lead to a full 3D articulatory model displaying both the inside and outside of a talking head, to be used in e.g. speech training/language teaching applications.

In the EU project PF-STAR, we aim at developing the extra-linguistic capabilities of the talking head. We concentrate on realisations and evaluation of the visual aspects of emotions and interaction/communicative signals, useful in e.g. conversational spoken dialogue systems.

In the EU project Synface, we work together with the Hearing group in the department and groups from England and Holland to develop and evaluate a system using our talking head that can help hard-of-hearing persons in telecommunication. The Babel Infovox company

(now part of Acapela Group) is the industrial partner.



The Synface telephone prototype

Speech technology and disabilities

Development and application of speech and language technology for disabled persons has a long tradition in our department. Apart from the Synface project, mentioned above we are involved in several other EU projects and projects with CTT partners. In the EU OLP project, where we co-operate with the Hearing Technology group, we investigate the use of speech recognition/analysis techniques for speech training of hard-of-hearing children.

Speech and language technology for motorically disabled and non-vocal persons is another area of interest. Research on communication disability has been designated a priority area at KTH. A large national project aiming at computer support programs for persons with reading and writing difficulties has supported part of this work. Our part of the project was concerned with text prediction. Currently we are the Swedish node in the EU project WWAAC concerned with symbol communication.

For an extended summary of external activities and projects, see page 23, *National and International Contacts*.

Open source software

The open source software developed in the speech group has been downloaded by many sites.

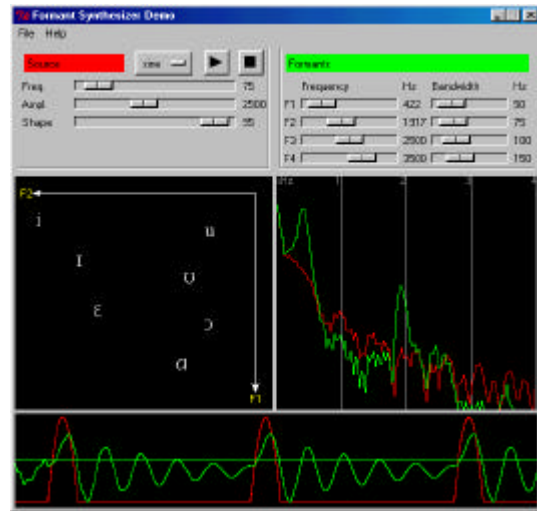
[//www.speech.kth.se/speech/speech_software.html](http://www.speech.kth.se/speech/speech_software.html)

Snack is an extension to the Tcl/Tk scripting language that adds commands sound I/O and sound visualization, e.g. waveforms and spectro-

grams. Snack serves as a general audio platform giving uniform access to the audio hardware on a number of systems.

Many applications have been created through Snack, including a general speech analysis and synthesis facility WaveSurfer and a re-implementation of the classical OVE 1 vowel synthesiser.

The popular ESPS Waves software is not on the market any longer. Through a donation of rights from Microsoft and AT&T of that software we have now made program modules available on our website, and have included part of the functionalities in current releases of WaveSurfer.



The classic OVE 1 re-implemented in Snack – available as open source.