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.

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 articulatory modelling, multimodal dialogue systems and natural language processing.

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 year 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.

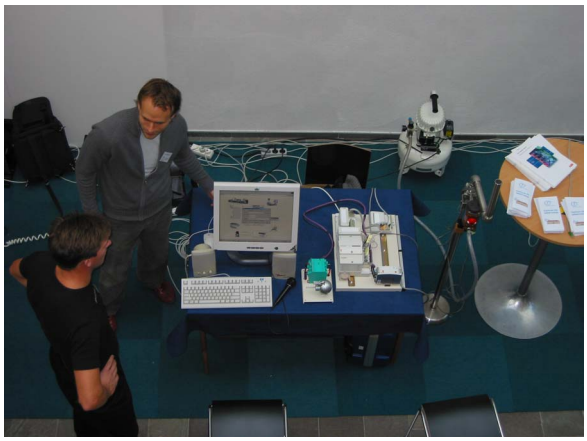
With the limitations of current speech technologies, both for recognition and understanding and for speech generation, the interest in "real" systems has led to an increased awareness of the problems raised by system errors, especially in recognizing user input, and the consequent confusion that such 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 initiated during the year along these lines. The new Higgins project will specially focus on error handling and some WOZ-experiments have already been conducted. 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 AdApt user interface with the animated agent Urban.

The Higgins user interface used during the WOZ-experiments.

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. Two applications have so far been addressed. A first baseline system has been built in an industrial environment in which a commercial platform developed by ABB is integrated with the PipeBeach Voice Web product. The second system concerns a reception application described in the section *Speaker characteristics* below.

In the EU project MultiSense we have started to implement a spoken dialogue system for a medical application.



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 during the year 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 of speech in terms of boundaries and groupings. The modelling includes different communicative situations and is based on 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.

Speech and language databases

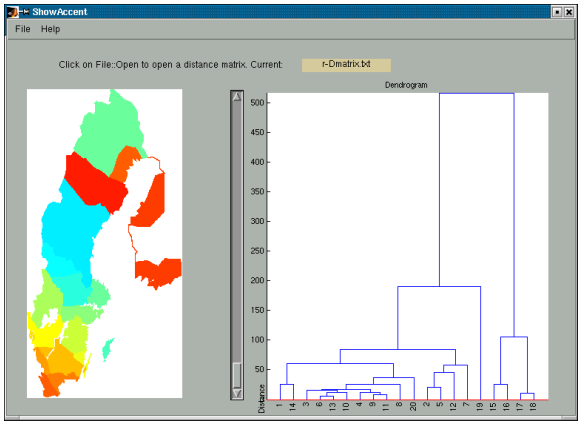
We see an expanding interest in studies on speaker variability, especially in the context of speaker independent/speaker adaptive recognition. Large text corpora are increasingly important for language technology developments. We have participated in several large efforts to build telephone speech databases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

Recording the SpeeCon database in the living room condition.

We also developed several databases primarily intended for speaker verification research. Large text corpora have been collected, containing 150 million words for use in e.g. language model experiments.

A database has been recorded in co-operation with the CTT partner Telia Research. It combines sound and video recordings with 3D registration of articulatory significant points on the face. It contains 1.5 hours of read speech from one speaker. It is primarily intended for our multimodal synthesis development.

Our tool for automatic segmentation of speech has been improved. On the TIMIT speech database we have achieved 90.6% correct segmentation (within 20 ms of the manual labels). A new speech recogniser able to handle large vocabularies is under development. It is based on Finite State Transducers (FSTs). This makes 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. A fast phonetic recogniser based on Artificial Neural Networks has been developed within the Synface project. Regarding robust recognition we have shown that a rather simple method for noise compensation favourably competes with what is available in commercial recognisers. In a thesis report, a thorough analysis has been made of the possible use of speech recognition for Bilprovningen (the official Swedish car inspection body). A demonstrator application was also built using the CTT Toolbox ATLAS. Another thesis project studies the use of dialectal information for speech recognition in the SpeechDat database. A result of this can be seen in the figure below.

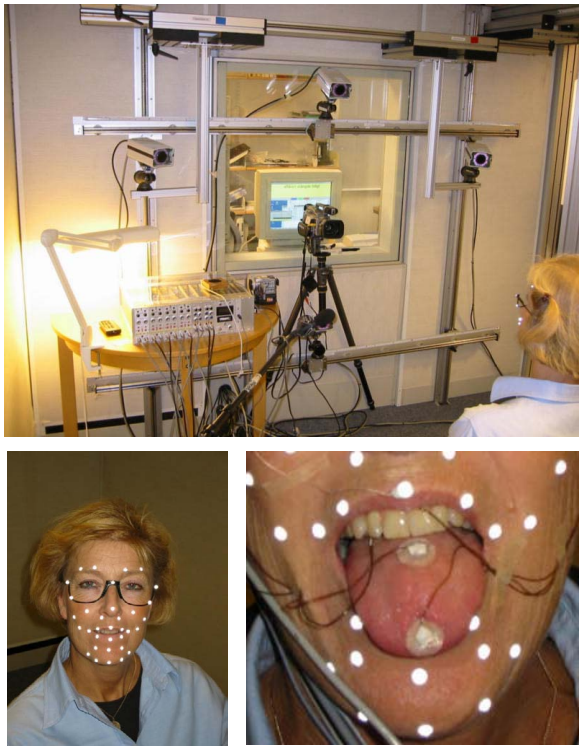


Swedish R-sound distribution. On the right is a dendogram that displays a "phonetic" distance between 20 different dialectal variants based on parameters used for speech recognition.

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. Articulatory models have recently attracted interest in our laboratory. 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.



Simultaneous recording of internal and external articulation

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-sagital EMA data.

Speaker characteristics

A system for text independent speaker verification has been developed. During the spring, we participated in the yearly international evaluation workshop NIST together with around twenty other systems from eleven different countries. The result of our system was positive considering the short development time. We had useful experience and inspiration from the evaluation. In the speaker verification domain, we are also engaged in the European COST 275 project.

The "PER" project is an effort to build an automated entrance receptionist, PER (Prototype Entrance Receptionist). It operates in the central entrance to the department. The purpose is to create and experiment with alternative speech based means of controlling access to the premises for employees and occasional visitors.

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.

A speaker adaptation service (TillTalad) has been designed to be independent of any specific application. A user who wants to adapt models to his/her voice, calls this service and records a number of adaptation sentences. The produced adapted phone models are then downloadable over the Internet to any application. This procedure reduces the adaptation effort for the user as well as for the service provider.

Research on discriminating between speech and music has resulted in a reliable technique that uses differences in temporal structure and spectral properties.

Tools for education and prototyping

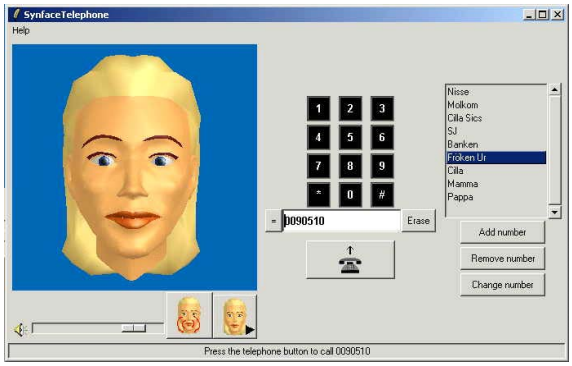
Our work on new tools continues. It has resulted in a new set of student labs in speech technology. An interactive dialogue system was created in which students can change and expand the system functions. A new framework for speech synthesis is the topic of another lab. These labs have been used and evaluated in several classes since 1999. This and other software developments at the department have changed the working environment for many projects. Fast prototyping based on modules is now part of general experimental designs.

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 tele-communication.



The Synface telephone prototype

Speech technology and disabilities

Speech and language technology for motorically disabled and non-vocal persons is a major research area. Research on communication disability has been designated a priority area at KTH. Several ways of increasing the communication speed have been investigated including interactive text prediction based on linguistic principles. 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 27 on *National and International Contacts*.

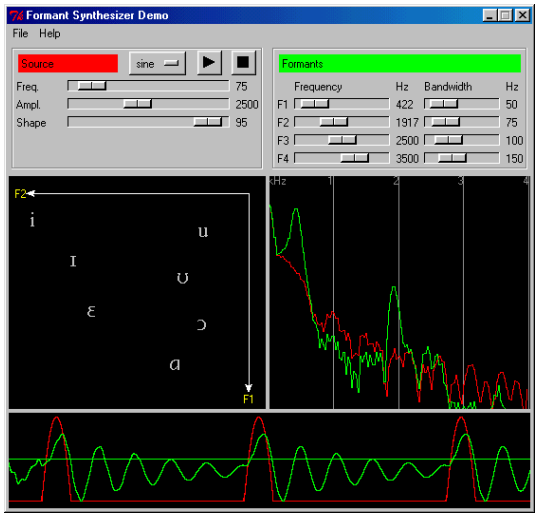
Open source software

The open source software developed in the speech group has been downloaded by many sites.
//www.speech.kth.se/speech/speech_software.html

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