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 has now expanded to about 35 researchers and research students, a few of them working part-time. The group includes CTT, the *Centre for Speech Technology,* that was established in 1996. The second phase that started in July 1998 was favourably evaluated in the the spring of 2001, resulting in a continued support of the centre in its third phase starting July 1, 2001. The organisation of CTT is presented on page 9.

The work in the speech group, including CTT, covers a wide variety of topics, ranging from detailed theoretical development of speech production models through phonetic analyses to practical applications of speech technology.

EuroSpeech 2001

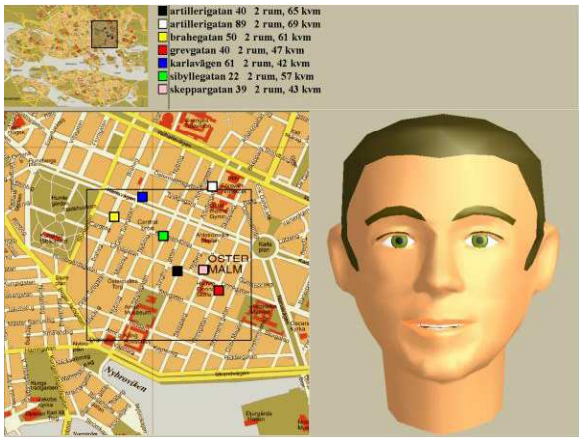
Together with colleagues in the Nordic countries we organized EuroSpeech 2001, held in Aalborg, Denmark, September 3-7, 2001. The conference was named "EuroSpeech 2001 – Scandinavia – An Interspeech Event", to signify the close co-operation that now exist between the Eurospeech and ICSLP series of conferences. The conference had a few new features. One was the full paper submission and another the ESE, Eurospeech Special Events, that were run during the whole conference devoted to diverse topics, not normally covered in the regular paper or poster sessions. (http://eurospeech2001.org/). The speech group was in the "top ten" group of accepted contributions to the conference.

Spoken dialogue/AdApt

A main focus of CTT is research on multimodal dialog systems. The motivation 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.

Building on experiences from the August system we have now implemented AdApt, a multimodal dialogue system for discussing and evaluating apartments for sale in Stockholm. The real-estate domain is particularly interesting from a research point of view. An apartment is a complex object that has properties that are suited for graphical presentation (e.g. its location in the city) as well as for presentation through spoken language (price, description of interior details, etc). In addition to spoken input, users have the possibility of providing the system with graphical information, e.g. referring to apartments by clicking their icons or marking areas on an interactive map of Stockholm. One of the areas studied in the project comprises various aspects of human-computer interaction in conversational multimodal dialogue systems.

The AdApt project was selected for display at the i3 Research Village at Comdex in Basel, September 2001 and has been presented at different national and international scientific meetings.



The AdApt user interface with the animated agent Urban.

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 a large effort to build telephone speech databases - the EU SpeechDat-project. This corpus consists of telephone speech from 6000 speakers. The databases are now available through ELRA. The August system was used to collect a database of more than 10,000 spontaneous utterances, which included a number of recordings of child voices. In the present EU project SpeeCon we are responsible for collecting the multi-microphone Swedish database recording in different environments. The database consists of material from 45 minutes 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 hour read speech from one speaker. It is primarily intended for our multimodal synthesis development.

Speech recognition

New acoustic models have been trained for speech recognition on the large SpeechDat database. We have shown that phone models trained on the task-independent, large speech database can provide high recognition accuracy in a certain application if they are adapted using a (much smaller) task-specific corpus. This procedure dramatically reduces the effort involved in creating new applications, since the required size of the adaptation speech data is much smaller. Experiments using glottal source adaptation have shown very promising results. Especially interesting is the work on phonetic categorisation feature-based context information. HMM-methods have been used in comparative experiments on speech signal representations. Special effort has been devoted to model speaker consistency within an utterance. Presently a Java-based recogniser, Ace, is being developed. Besides portability the main feature is the possibility of integrated acoustic and language modelling

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. 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. Special emphasis has been placed on analysis of female voice. Generalised observations from the analysis work are now implemented as rules for speech synthesis.

Different measurements for the 3D articulatory model (EPG, EMMA and MRI)

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.

PER (Prototype Entrance Receptionist) controls entrance to the department

Speaker characteristics

In the context of speaker verification we are engaged in the European COST 275 project. Several studies have been performed concerning human and technical impostors in speaker verification.

CTT Bank is an experimental demonstrator that will investigate the potential advantages of using speech for identity verification and user commands in bank telephony. CTT-bank has been evaluated outside CTT as part of a MSc thesis.

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 and emotions are studied. The long-lasting efforts on improved prosodic models and segmental synthesis continues.

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.

In the Teleface project, we work together with the Hearing group, investigating the usefulness of synthesised faces for hard-of-hearing persons in telecommunication. This project received financing through KFB, VINNOVA and Hjälpmedelsinstitutet, one of the CTT partners. The project creates several challenges – how can articulation be extracted from the speech signal and how shall the face animation be optimised in quality and time? The teleface concept forms the base of an EU project Synface, that started in 2001 with participation from England, Holland and Sweden. The project aims at developing and evaluating a prototype for the three languages.

The Teleface/Synface project was selected for display at the i3 Research Village at Comdex in Basel, September 2001.

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 25 on *National and International Contacts*.

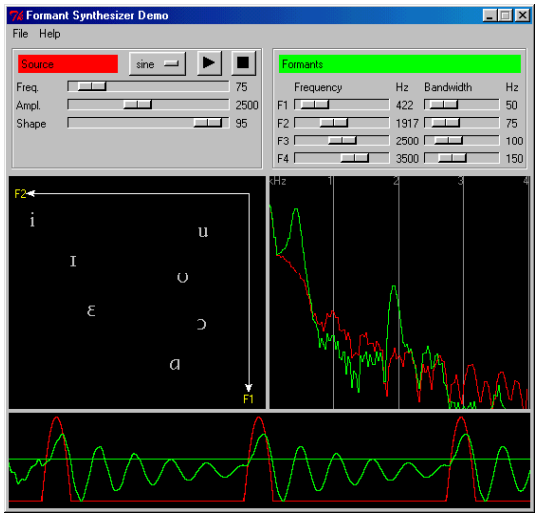
Open source software

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 & 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 we have now full access to that software, with the intension to include part of it in future releases of WaveSurfer.



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

Filename: speech2001.doc
Directory: I:\annual\Annual2001
Template: \\WINTER\install_off97\template\tmh\qpsr.dot
Title: Fonetik 96
Subject:
Author: cathrin
Keywords:
Comments:
Creation Date: 2002-05-02 16:28
Change Number: 6
Last Saved On: 2002-05-22 16:13
Last Saved By: Personal
Total Editing Time: 34 Minutes
Last Printed On: 2002-06-12 10:38
As of Last Complete Printing
Number of Pages: 4
Number of Words: 1 563 (approx.)
Number of Characters: 8 911 (approx.)