Another half a century in speech research*

Rolf Carlson and Björn Granström

In this paper, inspired by Gunnar Fant's "Half a Century in Phonetics and Speech research", we want to make a personal review of our own professional life. We were, somewhat violently, struck by the fact that we now also can look back on half a century in the area. It all started in 1969, just in time to be part of Gunnar Fant´s 50 year celebration. This means that 2019 is also the year to celebrate 100 years since Gunnar was born.

We had both been studying electrical engineering at KTH since 1965. The last semester we took Gunnar´s course, Talöverföring, which literally translates to Speech Transmission. We did it mostly "for fun". At that time we actually did not need the credits for our exam, but wanted to study something more human related, to balance the mainly technical, machine oriented topics. Talöverföring contained so much more than transmission.

After a very inspiring course and a successful exam, we dared to approach Gunnar with the simple question: "Where should we go if we want to engage in the technical speech communication area". And Gunnar gave the unexpected, but equally simple answer: "Start work with us". One deciding factor in this fortunate meeting might have been that we brought some spectrograms that Gunnar had distributed in the class as material to practise "spectrogram reading", i.e. to transcribe a spectrogram without access to text or sound. And Gunnar was obviously content with our results.

In retrospect we understand that we asked our question at a very fortunate moment. Gunnar had just "lost" two of his co-workers and earlier students. Sven Öhman had moved to Uppsala University for a chair in phonetics and Björn Lindblom had a position at Stockholm University where he in 1973 became professor in Speech Physiology and Speech Perception.

When we started, the department was in a strong growing phase and most new young scientists were recruited from electrical engineering at KTH. Two more of our KTH classmates, Inger Karlsson and Lennart Nord started as MSc thesis students, and were then employed. And the next year Kjell Elenius and Mats Blomberg followed the same track.

We have been fortunate to work in an academic area that has developed tremendously during the last half century considering theory, tools and applications. Before World War II phonetics was still to a great extent relying on auditory observations and introspective registrations of articulation, but after the war things had started to change. Analogue equipment for speech measurement and synthesis had started to emerge.

---

*This paper was first written in 2019, but due to diverse factors is now presented at the first post-pandemic, non-hybrid Fonetik 2022 conference

To develop these devices, competence in electrical engineering was paramount. Gunnar Fant's laboratory at KTH was one example of the pioneering labs. The sound spectrograph, a quite expensive apparatus, had spread to many laboratories, and the first speech synthesizers had appeared, needing many cubic feet of complicated electronics and often also in need of frequent maintenance. We started at an interesting moment when digital technology slowly started to be available - but more about that later.

It should be understood, that this is our very personal account of the mostly early part of the last 50 years and it is not aimed at a fair and balanced description of the development of speech science internationally or not even at the lab. We are mostly leaving the Music and Hearing part of the Department of Speech, Music and Hearing aside.

The beginning of our doctoral studies

Compared to the very well scheduled and regulated doctoral studies that is the rule presently we had a very different experience. When we started at the lab it was more as junior assistants in different projects – and at an early stage we were involved in the only undergraduate course in the department – Talöverföring (Speech Transmission), where we soon got the full responsibility for running the class. At that time there were no scheduled doctoral studies. Very few of the senior researchers, except Björn Lindblom and Sven Öhman, had a pronounced goal of becoming doctors. As doctoral students it took us almost eight years to graduate in 1977 – at that time a speed record for Gunnar's students in Speech Communication. Many of the internationally recognised senior researchers in the department actually graduated much later than we did. No formal teaching on the PhD level existed. Rather a very efficient "learning by doing" strategy was practised. Of course, reading Gunnar´s basic books, like Acoustic Theory of Speech Production, had high priority. Björn Lindblom inspired us to widen our technical training by taking (undergraduate) classes in Phonetics, General Linguistics and Pedagogy. The latter was partially motivated by the hope to be better prepared for running the Talöverföring course at KTH. Totally we earned points at the university corresponding to a BA. When we later got responsibility for our own doctoral students, we advised them to take at least the introductory course in Linguistics – and arranged for credits when doctoral studies at KTH asked for more structure. The close interaction with different faculties must be regarded as a key component in the development of our department. From the start it seemed natural to us since we are dealing with the speech communication process, a topic that is inherently apt to a multidisciplinary approach. Talking to colleagues in speech communication and technology we have understood that this is far from globally accepted.

We have been lucky to "grow up" in a scientific community that encouraged interdisciplinary exchange. This is much due to a very understanding view of our grant agencies. Already in the 70ies Forskningsrådsnämnden (a cooperation body for the agencies) financed the Speech, Sound and Hearing program. It organized summer schools, where essentially all research students in our area – from arts, science and medicine – met with many senior researchers. When we reached a level when we could apply for our own grants, a new initiative had just been launched – the Swedish Language Technology Program, stimulating continued

cooperation with researchers across Sweden. Another initiative was the sequence of national research symposia in phonetics for doctoral students. But, when we and our fellow students started to graduate, we all missed this enjoyable opportunity to meet. Finally, in 1986 the series of Svenska Fonetikmötet (the Swedish phonetics conferences) started. Actually, with some international/Nordic participation. The conference was much seen as a way to meet and discuss in a friendly atmosphere - often the first occasion for new students to share their research results. To make it easy the conference language was Swedish. Today the conference is mostly in English – today probably the easiest language also for many Swedish researchers when they want to communicate their research. The annual Svenska Fonetikmötet continues and has since 1996 been supported by Fonetikstiftelsen (the Phonetics Foundation), based on the revenues from ICPhS 1995 in Stockholm. Most of the Swedish phonetics and speech science community contributed to the congress without pay – hence the substantial revenue.

Early days

Working in an internationally recognised department with an extensive professional network meant that we already at an early stage had the opportunity to make interesting study and conference trips. We both grew up during a time when the Interrail pass was still not invented and unlike most of the kids of today we did not have the experience of travelling abroad outside Europe. Our first conference trip was to Berlin during our first spring at the lab. We were going there for a cybernetics conference where Gunnar was an invited speaker. We were travelling there by train – at that time the argument for train travel was more economy than saving the planet. But the travel itself was exciting, especially since we had to go through East Germany and pass the well-known Check Point Charlie by foot (On our travel back we actually managed to miss the train in East Berlin due to problems with the connections from West Berlin.) The plan was that Gunnar should come later by air. However, that did not happen, for reasons that are still not clear. This was admittedly long before cell phones and email. And somehow Gunnar´s absence was not communicated to the organizers, so we had to announce it in a plenary meeting. Thus, our first appearance at an international conference was not even in speech science. One of the more odd interactions during the conference was in a breakfast meeting where one participant wanted to hear our views on being in the first generation that did not have to die, thanks to the progress in medicine.

The next trip took us to Great Britain together with Lennart Nord, one of our classmates during our MSc studies at KTH. This time it was a ferry and car trip, which was very convenient.

Taking the ”ferry” to Newcastle. Definitely not ro-ro.

Our extensive trip took us from England to Scotland and to many of the well-known labs, like UCL in London, JSRU in Malvern and of course to Edinburgh University. The trip was very inspiring and as Gunnar´s students we were extremely well received – and taken care of. James Anthony, in Edinburgh, who developed the OVE II “competitor” PAT together with Walter Lawrence, even offered us to borrow a cottage in the highlands, which we gladly accepted.

Lennart Nord, Rolf Carlson and Björn Granström in James Anthony´s highland cottage.

Combining work and pleasure in this way was encouraged by Gunnar and there was, as far as we ever experienced, no administrative red tape that prohibited this – as long as it did not add to the cost for the lab.

At that time speech communication research, to us, appeared to be a very small community that operated as a big extended happy family. Long before we started Gunnar had collected many of his colleagues to the first SCS - Speech

Communication Seminar - in 1962, with less than 100 participants. In a sense this was a precursor of the international meetings like Eurospeech and ICSLP that much later merged into the present day Interspeech with more than 2000 participants. In 1974 we had the opportunity to help organizing the second SCS, also in Stockholm, now with several hundred participants. The attendance list contains many, if not all, of the early pioneers of our science.

At this time no external professional conference organizer was involved. Everything was in-house. There is possibly no better team-building activity. At that time simple things like printing legible, large size name tags could not be done by our department printers. But we both had taken a class at KTH in photography and were quite familiar with a dark room and produced nice “large font” typewritten name tags. Before the conference the Formant Orchestra was formed and during the conference banquet at the Opera Restaurant, Gunnar together with Inger Karlsson made a performance of Swedish folkdance accompanied by the orchestra.

That we were at all able to organize SCS without external conference organizers was of course due to the structure of the very self-contained department at that time. Today researchers and students are used to do everything themselves with efficient computer support, and maybe backing off unusual tasks to a more central administration. Forty years ago, publication, communication, economy and scientific experiments all needed their specialized support personnel. In the annual laboratory reports from that time we count around 10 persons engaged in this kind of support for around 30 academics, including researcher, teachers and doctoral students.

The department had a workshop and highly skilled hardware developers and engineers. Most analyses, such as fundamental frequency or long time amplitude average LTAS, were performed with specially designed modules and some of them could also be connected to the computer. In the 80ies the computer development practically exploded and computers with formal coding languages made life so easy also for research. Speech analysis and signal processing were just one click away and soon the personal computers were on our tables.

Most of the publications from this time are available in the department scientific report STL-QPSR, still available on our web site, (http://www.speech.kth.se/qpsr/) due to a heroic effort by our colleague Giampiero Salvi. A special edition was created 1979 to honour Gunnar’s 60th birthday in which all members of the department contributed about their views and activities.

A fully equipped mechanical workshop, used by both scientists and technicians was very useful before the age of computer simulation. In the picture Frans Franson (left) and Eric Jansson, both in the music acoustics group.

Our librarian, Surjadi Gorda, in front of the huge preprint collection from all over the world. Preprints were to a great extent received in exchange for our globally distributed quarterly report QPSR.

With the risk of being looked upon as dinosaurs we miss that situation, with daily interactions with the administrative and technical staff. It created the more varied working environment that constituted Gunnar's professional family.

The mix of personnel made it possible for us to navigate through the KTH organization more like a research institute without too much interference from the KTH hierarchy. If we made good research and attracted enough money we could make most of the decisions ourselves. An environment quite far from current state of affairs. This situation also tightened the group and formed a unique atmosphere of creativity accompanied by pleasant parties, celebrations at the top floor of Wennergren Centre and other activities outside KTH. The tradition that everybody should contribute to lab activities such as the Christmas party is something that has stayed on during all these years and it has been an important way to tighten the group and make it special.

Garden party after Rolf´s and Björn´s dual thesis defence on May 17, 1977.

Garden party after Rolf´s and Björn´s dual thesis defence on May 17, 1977.

International associations and meetings

Our choice of speech communication and technology as our academic field was very lucky and not so much due to a thorough and well considered search for possibilities, but rather the fortunate participation in Gunnar Fant´s inspiring course. We stumbled into an area that was under strong development and very internationally oriented. There were no international periodic meetings dedicated to our area – even if Gunnar had organised two Speech communication seminars with a 12-year interval. After 1974 most of us did not want to wait for 12 more years for the next conference, but presented our research to other more general

conferences, like ICASSP and ASA. Starting in Europe the need for an organisation dedicated to Speech communication and technology emerged. We were invited to join this effort and Björn took a place in the first board of ESCA in 1988. As an association we started a number of things useful for the speech community, like conferences (Eurospeech, the first one organized in 1989), newsletters and topical tutorial research workshops (ETRW). Björn was given the responsibility to manage the workshops. Topics contained both established and emerging areas. Like the first STiLL workshop in 1998 (Speech Technology in Language Learning, later renamed to SLATE) that we organized together with Anne-Marie Öster in the Stockholm archipelago, a good example of combining work and pleasure. Many of the workshops also resulted in the formation of ESCA special interest groups (SIGs).

Hiroya Fujisaki at the same time started to plan for ICSLP (the International Conference on Spoken Language Processing) with the first meeting in 1990 in Japan. For a long time the two conferences were held on alternating years. When Björn´s term of duty ran out in 1997 Rolf joined the ESCA board with the same duty and the work also continued to merge Eurospeech and ICSLP into the annual Interspeech, backed by the international association ISCA, International Speech Communication Association, replacing ESCA. Areas like dialog modelling successively became important subjects for the progress in Speech Communication and required new dedicated workshops like the Error Handling in Spoken Dialogue Systems workshop that was organized in Switzerland 2003 together with Gabriel Skanze, Julia Hirschberg and Marc Swerts.

In 2005 David House became the third representative of the lab on the ISCA board.

In speech research and technology we obviously needed research and tools dealing with written language. For a long time the reverse did not appear obvious, but the discipline (text only) was often referred to as “natural language processing”. What could be more natural than speech communication? In an effort to ease the contact between disciplines, Elsnet (the European Language and Speech Network) was created. Björn was on the board for many years and was active in promoting the popular Elsnet summer schools. One example was one that we organized at KTH in 1999, Milass (Multi-modality in Language and Speech Systems), the first of its kind.

On the Nordic scene cooperation of a similar kind took place. With financing from NordForsk (the Nordic Research Board) a project, VISPP (Variation in Speech Perception and Production), was created with the aim of increasing Nordic cooperation. Also, in this cooperation the organization of summer schools was one of the more important tasks. The series of summer schools had a wide participation, also from outside the Nordic region.

Long before Sweden became a part of EU in 1995, European research projects have played an important role for our group. Besides the great scientific value and the international networks they supplied, they also contributed to significant economic support that was hard to find in Sweden. Only between 1996 and 2006, during the CTT period (described below), we were part of 17 European projects. One challenge in these large-scale projects is the amount of coordination needed. Detailed reports

and other requirements that such projects require was sometimes regarded as prohibitive, but we felt that it was well worth it. Inger Karlsson became an expert to handle these demands and her knowledge was appreciated by several EU projects at KTH, also outside the lab.

Several projects were focused on building speech corpora for different needs. As the research successively focused on data-driven methods, large language and speech corpora were increasingly important. Kjell Elenius played an important role to organize and guide the collection, labelling and structuring of all data in many projects like SpeechDat.

Our early projects

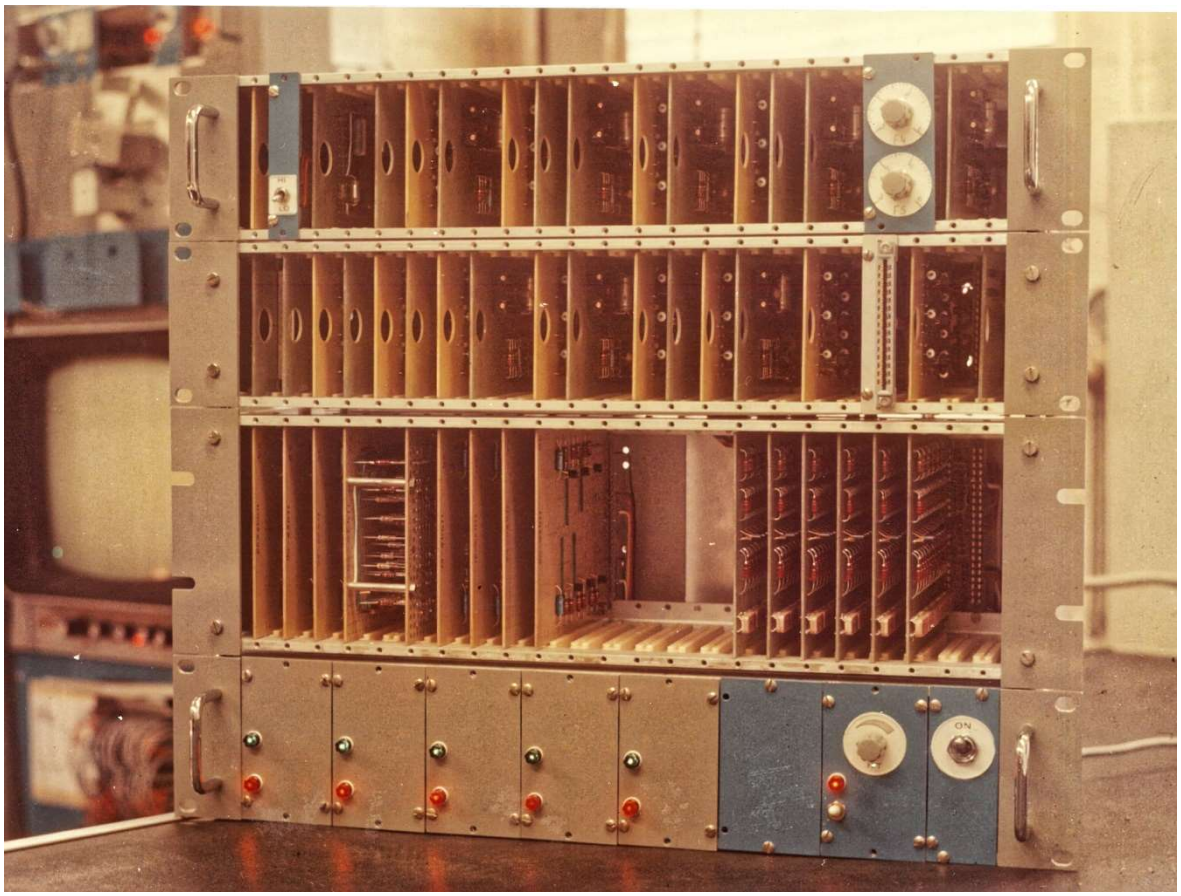
The first projects we were involved in concerned speech perception. The very first study was one of Gunnar Fant's "babies", the two-formant approximation of vowels. With the OVE II synthesizer the two lowest formants could be dynamically controlled and it seemed quite sufficient to produce acceptable vowels. But from speech analysis we know that two vowels with the same F1 and F2 could sound very different. A striking example is the unrounded and rounded Swedish vowels /i/ and /y/. Gunnar´s question was: *Could vowels still be reasonably approximated by two formants by changing F2 to a different position called F2'.*

We created a test environment where digitally synthesized vowels with four formants could be interactively matched to a two-formant vowel. The short answer was: *Yes, indeed!*

Frequently new research results lead to new questions. How did it happen? Could auditory models explain our results? To better understand this, we created one of our own – that we named the Domin model. Basically, it was a filter bank model of the cochlea followed by a zero-crossing count in each filter. F2' was well approximated by the dominant zero-crossing frequency for the full vowel. Was this then a peripheral cochlear explanation? Possibly not. By splitting the formants between ears in a dichotic experiment (e.g. F1 and F3 in the left ear and F2 and F4 in the right ear) we still perceived the same vowel.

But new challenges were ahead. Many perceptual studies at this time was made with "tape splicing", where speech segments on a physical audio tape were cut out and glued together with segments from a different recording. Placing the cut at the intended place was not trivial and keeping track of all the tape fragments was a challenge. Since we had access to the laboratory computer, the obvious solution was to do it digitally and the Mix program was created. It was later expanded to include multichannel analysis for physiological experiments and spectrographic analysis. Many studies in the department was made with this tool, including perception experiments of segmental consonant cues and studies on durational perception.

Gradually we focussed more on speech synthesis, first as a research tool and later as a system for diverse applications. It was obvious from the pioneering work at KTH, Haskins, MIT, Bell Labs, JSRU, Edinburgh and other labs that a formant synthesizer, appropriately controlled, could produce intelligible speech. At KTH Johan Liljencrants had designed a digitally controlled version of the OVE speech synthesizer called OVE III that he connected to our laboratory computer.



The computer controlled OVE III designed by Johan Liljencrants

The big new challenge in the early 70ies was text-to-speech. A reading machine for the blind was one visionary example, were the new technologies of optical character recognition and speech synthesis could be combined. In our linguistic studies at Stockholm University we came across the new generative grammar and we were especially inspired by Chomsky and Halle's "Sound pattern of English". For persons with an engineering training it was amazing and stimulating that language processes could be described by features, algorithms and formulas.

But obviously this was not so easy to program. The generative phonology used context sensitive rewrite rules, but what we wanted to generate from text was continuous parameters for the synthesizer. So we started to implement a compiler based on the notation used in generative phonology, but expanded it to parameter manipulation. The symbols, features and parameters could be used both in the context and transformation part of the rules. Features could be positive, negative or undefined and parameters one or two dimensional. We named it RULSYS. It turned out that the description could be made very compact.

Today spelling-to-phonetic transcription is corpus based. At that time an exhaustive pronunciation dictionary was not available in digitized form, let alone that it could not fit the resources of our laboratory computer. For the initial systems we only included a list of function words, due to their irregular pronunciation and special prosodic function. Later when we copied a list of the most common 10 000 words in Swedish and created pronunciations for each of these words we were able to both improve the letter-to-sound rules and to

automatically generate an exception dictionary for the words not following our rules. Sets of rules could be combined into a complete text-to-speech system.

Due to the relatively simple development platform and a notation that was easy to use for non-programmers this evolved to the world's first multi-lingual text-to-speech system. The RULSYS approach in addition to our perception work was the base for our joint doctoral thesis. Also very rewardingly, it was the base for many cooperative projects and finally for a commercial exploitation – The Infovox system.

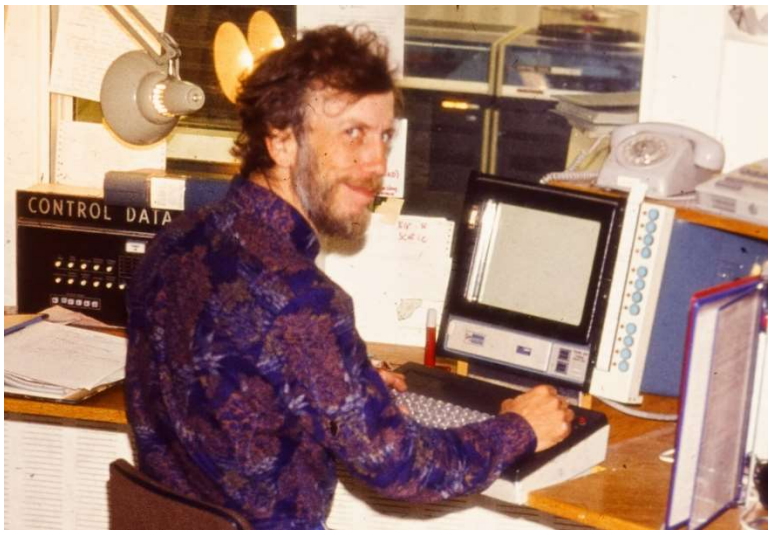
From punched cards to the cloud

The technical development has been extremely dramatic during our life at KTH and it has changed our life as speech researchers completely. It has been so exciting to be able to participate in this journey. 1969, when we started, also happens to be the year internet was born.

Many ideas were not possible to test at all until the time in the technical revolution was ripe. Looking back, perhaps we should not have tried so hard and spent so much time to overcome limitations in computer speed, memory and accuracy. But, on the other hand, how should it otherwise have been possible to push forward in the research to understand more of the speech communication process. It was certainly rewarding to hear the text-to-speech system speak and see the auditory model vibrate. And to see the systems come into use by functionally disabled users has been extremely satisfying.

We did our joint MSc thesis work on "modeling the transmission of electrocardiography signals in the body" to complete our education at KTH. At that time there was only one central KTH computer, which could be used for researchers, teachers and students. The software that was going to be run had to be coded in an ordered pile of punched cards. Each card corresponded to one line in the program. We delivered our software in a box to the reception of the department of numerical methods and picked up the printed result (or more often a bug report) the following day.

It was amazing to start working at TMH that had one of the first computers at KTH devoted to a specific research field. The CDC 1700 (Control Data Corporation) from 1966 was a 16-bit minicomputer with a 16 Kbyte core-storage memory. As the computer got older some of the transistor circuits got a little shaky and those cards were marked with a wire and sometimes it was helpful to knock on them a little. Johan Liljencrants, in charge of the machinery, had cleverly connected the data bus with an audio system, so it was possible to listen to the computer to check that it was running all right. The programs were coded using machine instructions and Kjell Elenius later made a heroic effort to write an assembler/dissembler program so we could take one small step up from the machine codes.



Johan Liljencrants, our computer guru, in full control of the CDC computer (early 1970ies)

This was the machinery that we used to make our first auditory models and the first text-to-speech rule compiler with corresponding runtime modules. A special challenge in this work was to implement a mathematical compiler and a floating-point calculator.

When the TTS system started to take shape, we successively needed more text to run through the system and to create a corpus to develop and evaluate the letter-to-sound rules. A Facit punch-tape machine was connected to the computer. And we managed to get punch-tape rolls from printing offices and newspapers in large cardboard boxes. When things went wrong, we rapidly filled the computer control room with serpentines of punched tape.

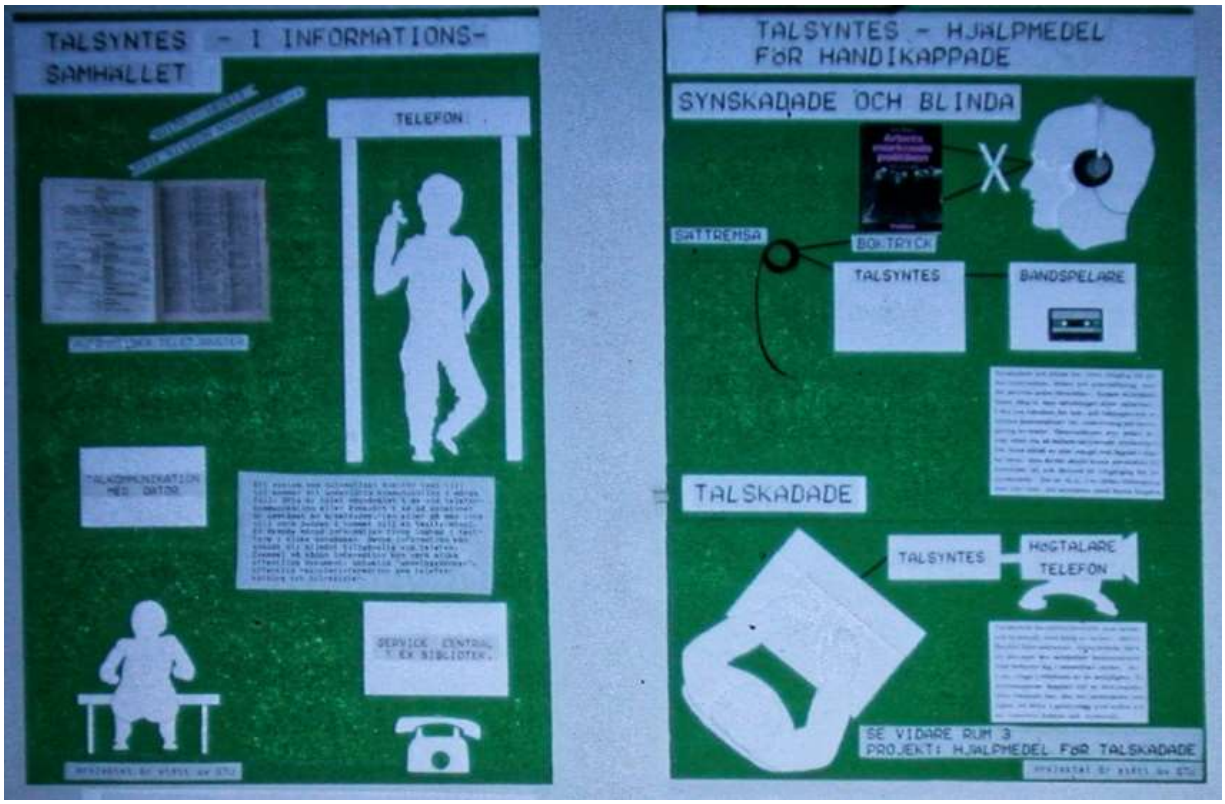
Rolf and Björn demonstrating the text-to-speech for the newspaper Expressen in 1976

The step was short to test and produce synthesized talking books and newspapers. Together with the Blind organization SRF and the Swedish Institute for the Handicapped we started a long-term cooperation to develop new talking aids. In a more advanced setting, we implemented an interactive system for blind users to read the newspaper Göteborgsposten, using the telephone network. This was done in cooperation with researchers from Chalmers. The first synthesis systems running outside our office were built using the Alpha computer in the late 70ies. Kjell Elenius made an efficient code converter for the CDC to Alpha transfer. The system had the size of a chest of drawers.

The first text-to-speech system we could try outside KTH (1977)

The first speech prosthesis in use

It was tested as a speech prosthesis for a boy with CP (cerebral palsy disorder). It was a touching and memorable moment to hear his “vocal” interaction with his parents for the first time in his life.



Outreach before poster printers. For one of the departments open houses big explanatory posters were built in Styrofoam. These two explaining applications of text-to-speech technology. To the left in a telephone reverse directory service and to the right for disabled persons.

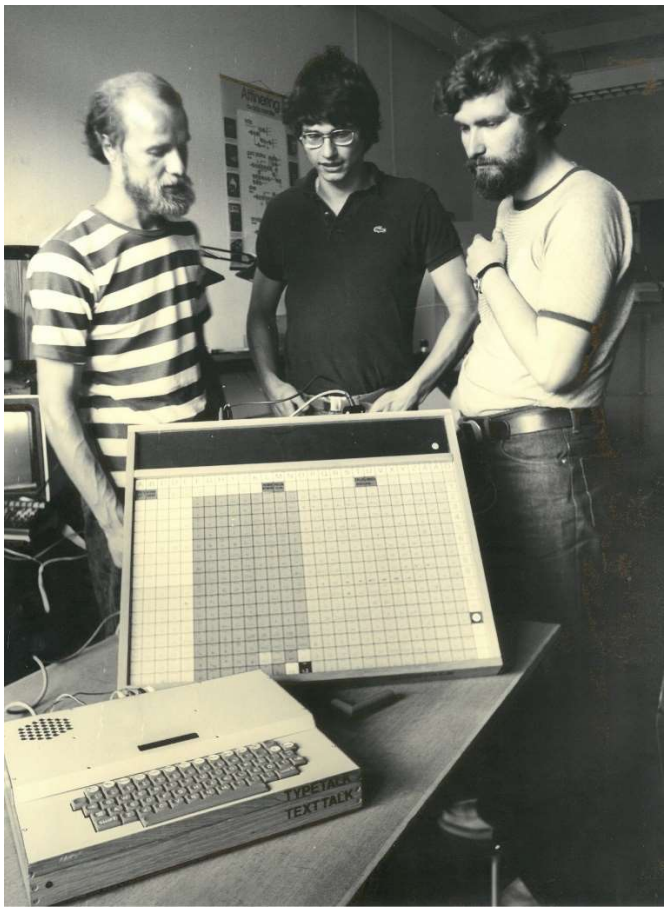
Taking research into public use, the creation of Infovox

The Motorola 68000 microprocessor, introduced in 1979, opened a fantastic new possibility to create a TTS system that was personal, affordable and accessible for all. We managed to get hold of one of the first developing hardware kits from Motorola and Kjell Elenius created software to translate the runtime module software of our TTS system to the microprocessor codes. But a fully functional computer based on the 68000 processor did not exist. So a computer was designed and built by Björn Larson and Lennart Neovius and a complete TTS system was finally running on the in-house built computer.

The synthesizer model was implemented on one of the first programmable signal processing chips - NEC7720. We got it directly from Japan together with a development system from a close contact and friend of Gunnar Fant at NEC in Japan. The challenge to build the first portable multilingual TTS system forced us to work day and night for a long period and when we finally could listen to the first sounds it was an unbelievable, memorable moment. This system was presented at the ICASSP meeting in Paris 1982. The multilingual system contained in a small

loudspeaker-looking box was presented as a poster presentation. Many visitors looked in vain for a computer behind the poster. We were immediately contacted by large companies like Canon and Olivetti. But we had other plans.

First international display of our portable Text-to-Speech system as a poster at ICASSP in Paris 1982. The "loudspeaker" box that Björn is holding contains the whole system. Rolf in the front.



Björn Granström, Lennart Neovius and Rolf Carlson and our first truly mobile TTS system and Blisstalk (1981)

We realized that now was the time to act and make our TTS accessible for the general public and, as a first step, for disabled users. Mats Blomberg and Kjell Elenius had implemented their speech recognition system on the same hardware and together we were ready to make the systems available on a commercial basis. The Infovox Company was created 1982 together with the governmental company SUAB, Svenska Utvecklingsaktiebolaget. Several demonstration systems were built. Some of these stand-alone boxes are still working and can be demonstrated at special occasions. After the PC revolution the text-to-speech hardware moved inside the PC as an extension card and finally the complete TTS system was transformed to a simple PC application.

We have already mentioned that the RULSYS system was flexible and language independent. And that we in 1982 presented a TTS and also a Bliss symbol-to-speech system, intended for persons with a communication handicap.

Thus, at the start of the Infovox Company we already had several languages implemented. The generic synthesis approach made it possible to cooperate with several researchers in Sweden and abroad, who were excited to structure their language knowledge into a strict and unforgiving rule structure. But we needed much more than a theoretical understanding of specific phonetic details. We were looking for context dependent phonetic realization rules, prosodic rules, letter-to-sound rules, digit pronunciation rules, simple parsing and much more.

A year after the creation of Infovox someone knocked on our door at KTH. It was a representative of a blind organization in the south of Sweden who came to express their gratitude of how the Infovox system had changed their life situation.

At a later stage Infovox was completely acquired by Telia, and in 2003 it was merged with Babel Technologies and Elan speech to form the Acapela group, but that is another story.

Uses and abuses of RULSYS

Our synthesis system, built on the expanded Chomskyan notation for context sensitive transformation rules, was primarily used for text-to-speech systems, but was flexible enough to be used in other applications. When we first started, we saw the tool as a productive way to experiment with knowledge-based models of speech.

Speech synthesis production experiments

A special feature of Rulsys was that it could use external variables in the rules that was used during the execution. These variables could for example be introduced at runtime by moving a cursor on the screen. We used this in many experiments where subjects could manipulate different aspects of the synthesis. Several studies on the temporal aspects of speech was conveniently done this way, where subject were asked to set their preferred duration of (groups of) segments.

Tone 1.5

In text-to-speech systems for Swedish (and Norwegian) it is sometimes hard to predict the correct word tone (Tone 1 and Tone 2) especially if you don´t have access to a large pronunciation dictionary – which we could not “afford” when we

started. No such dictionaries were freely available at that time. Using the wrong word tone could be at least confusing, giving rise to a dialectal impression and also affect intelligibility. The functional load on the tonal system is not paramount in Swedish. There are actually Swedish dialects (in Finland) where the distinction does not exist. But our target was standard Swedish. (Later on we actually run projects aiming at dialect synthesis). One idea that we wanted to evaluate was the cost of implementing a hybrid "Tone 1.5" between Tone 1 and Tone 2 and using that in ambiguous cases. The result might have been to reduce the effect of choosing the wrong tone, and still sounding reasonable. A similar "cost" argument was behind the next study, deaf synthesis.

Deaf synthesis

The speech of severely hard-of-hearing persons is often hard to understand. Much effort has been spent in teaching the persons a more understandable pronunciation, with varied success. A main difficulty is the lack of guiding feedback, but also a collective understanding on where to put the pedagogical effort. Some aspects of speech might be very hard to teach and may have marginal impact on the intelligibility/acceptability of the speech. The reverse is of course also true.

In cooperation with Anne-Marie Öster from the Hearing Technology group in our department we made an effort to demonstrate this. Pronunciations typical for the speech of severely hard of hearing persons were modelled in the text-to-speech system, like hyper nasalization and prosodic deviations. With variation of the simulated speech corresponding to different corrections, synthetic "deaf speech" could be produced and evaluated in listening tests. This was carried out as a feasibility study and did not get further funding. A typical example of the condition that in research we often need to "follow the money"

Database searches

In the 80ies large digital speech databases started to be part of our tool box. A problem was to conveniently search the database for interesting content. The RULSYS notation was used to point to examples of speech in the segmented and annotated database. Some measures could immediately be collected, like durational measurements and other found speech samples could be directed to further manual or automatic analysis, with minimal programming effort.

Creation of pronunciation dictionaries

As we pointed out earlier, large pronunciation dictionaries were not freely available when computer and memory technology made it feasible to include them in text-to-speech systems. As a sub-project within CTT we decided to create such a dictionary for Swedish. The simple and efficient procedure was to base it on a large (newspaper) corpus that was frequency sorted. The x most common words were passed through the text-to-speech system and the phonetic transcriptions were checked by a transcriber. The majority of the words were correctly transcribed by the system, and could be rapidly accepted by listening. We found out that both speed and precision was greatly improved by this procedure.

Name pronunciation – the Telia phone book

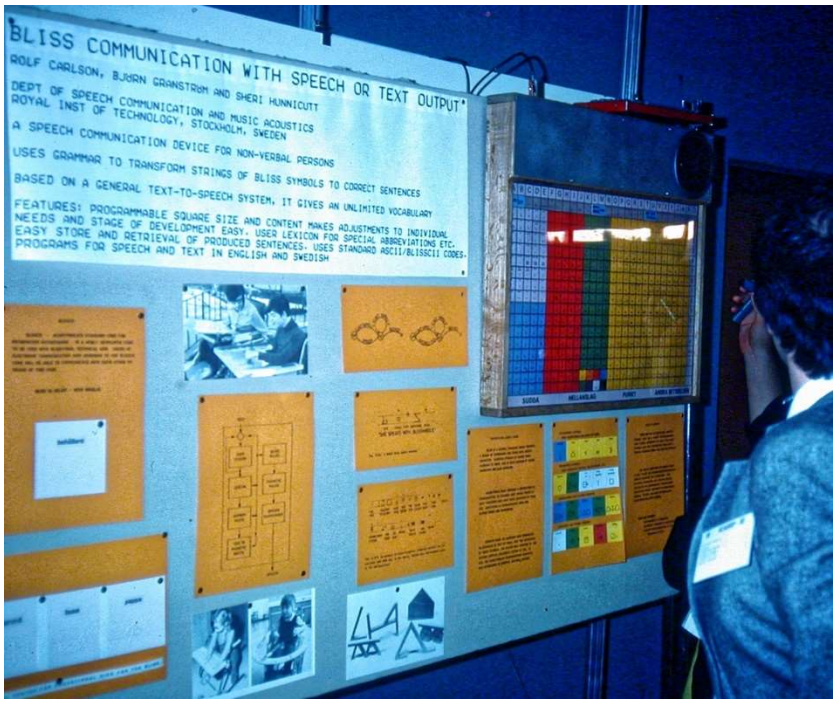
In an early EU-project, Onomastica, the objective was to create extensive name pronunciation dictionaries, as a base for "who has number xx", an automatic reverse directory service on the phone. Remember that internet at this time was not in common use and smart phones were hardly in sight. Our general text-to-speech rules were not so useful. Names tended to be exceptions to the rules. Much of the problem was that people used "misspellings" to make their name unique, which of course is much of the function of names. But this does not mean that the name pronunciation did not follow any rules. We could improve pronunciation by creating specialized rules, to minimize the transcription effort.

Bliss-to-speech

The Bliss communication system, created by Charles Bliss, a Canadian concerned with international communication. It was inspired by Chinese writing. A problem with Chinese for this purpose is of course that the meaning of the signs is not obvious from their appearance. Bliss created new symbols that should be easy to understand and learn. The system never got a universal acceptance, but one interesting use has been for person with communication disabilities, often combined with a motor handicap. A typical way of using the system was to select a sequence of symbols on a Bliss symbol chart. This was much faster than e.g. spelling. A problem we saw was that the person you "talked" with had to see the chart (and understand the symbols). Experiments in clinics in Canada used recorded speech and pronounced the base form of the symbols, one by one. We were intrigued by the possibility to feed the string of symbols to our synthesis system.

If we managed to make (context dependent) modifications to a string we should be able to generate something more useful, more similar to fluent speech. Sheri Hunnicutt put a lot of effort into modelling linguistic transformations with the help of RULSYS rules to generate well-formed utterances. We also made a prototype input devise based on the standard Bliss chart, were the symbols could be selected by passing a handle with a magnet above the symbols (see the photo). This device was also first presented at the ICASSP meeting in Paris in 1982. The system was developed for several languages and commercialized by a Swedish company.

Late night preparation for the ICASSP 1982 Bliss board poster presentation



The ICASSP 1982 BLISS poster

Dialect simulation

Sweden is geographically a large country, rather sparsely inhabited. This is probably a reason that several rather distinct dialects have developed. Around the turn of the century a large project SWEDIA 2000, involving three different phonetic departments, was created. The main objective was to record the dialects across Sweden (and part of Finland). The database has since then been used for different studies. One possibility that we studied in a cooperation with Gösta Bruce´s group in Lund was to use the data base for dialect synthesis that may have application in systems where regional character is preferred.

Literacy training

As part of "Centre mondial informatique et resource humaine" in Paris we suggested a novel and different use of a text-to-speech system. The background was that the literacy in native languages in Senegal, like Wolof and Fula was reported as quite low. By creating a text-to-speech system for theses languages, it should be possible for the students to explore and train the written language in a presumably efficient, interactive and interesting way. This should be possible to do without any demand on a technical infra structure. The regular correspondence between text and speech should make it easy to develop the system. The text-to-speech system was at this time, in the mid 1980ies, compact and low-power, easy to connect to portable solar power. The project was not realised due to, for us, unknown reasons.

Music realization rules

Human communication, by speech or by music has many similarities, often sharing the same instrument – the vocal organs. So it is no surprise that we also share many research objectives and have a lot of interaction. One similarity is in the written form. Neither text nor musical notation contain the full information for creating the acoustics of speech and music. The music group had designed a music synthesizer, Musse, and was interested in controlling Musse by rule. An obvious cooperation was to base the rules on RULSYS. Many phenomena could be given similar formulation, like slowing down in the end of some units like ritardando in music and phrase final lengthening in speech. Vocalise by Rachmaninoff was our first joint attempt together with Johan Sundberg and Lars Frydén and very fitting since it is sung with only one vowel. We used the synthesis in a fake application to a chorus position but were not accepted, due to the lack of consonants. Anders Friberg developed new methods to successfully carry this research direction further.

Early international travel

Going east

In 1973 Gunnar organized a meeting on speech perception in Leningrad together with his Russian friends Ludmilla Chistovich and Valery Kozhevnikov at the Pavlov Institute of Physiology of the Academy of Sciences of the USSR. In Björn's Peugeot

we drove from Helsinki on an adventurous journey. We presented our vowel studies for the first time to an international audience.

We are passing the Winter Palace in Leningrad in our car, 1973

This was our start for several trips to the Soviet Union, stretching all the way to the Mathematical Institute outside Novosibirsk, where we besides research discussions took part in the celebration of the October Revolution in freezing cold weather. It was a great experience to meet Russian researchers that often had problems to visit the western countries and to present their work. One fascinating project that we saw was a computer controlled mechanical synthesizer in the van Kempelen tradition. In general, we found excellent theoretical research on human interaction during our trips though the Soviet Union.

Some years later, in 1980, we set off by train to Beijing by the Trans-Siberian railroad. This we made for curiosity, rather than to save the earth from too much CO_2 emission. We shared the wagon with the Chinese ambassador from Austria and his wife who cooked their food on the minute stove available in each carriage. In the restaurant car the supply was running short when we approached the Chinese border. In the end the only items on the menu was rye bread (stored under the seats), Russian caviar and Russian Champagne! At the boarder the train carriages were lifted up and the bogies we changed to a wider gauge and a Chinese restaurant car was added. It was immediately completely filled by Chinese customers.

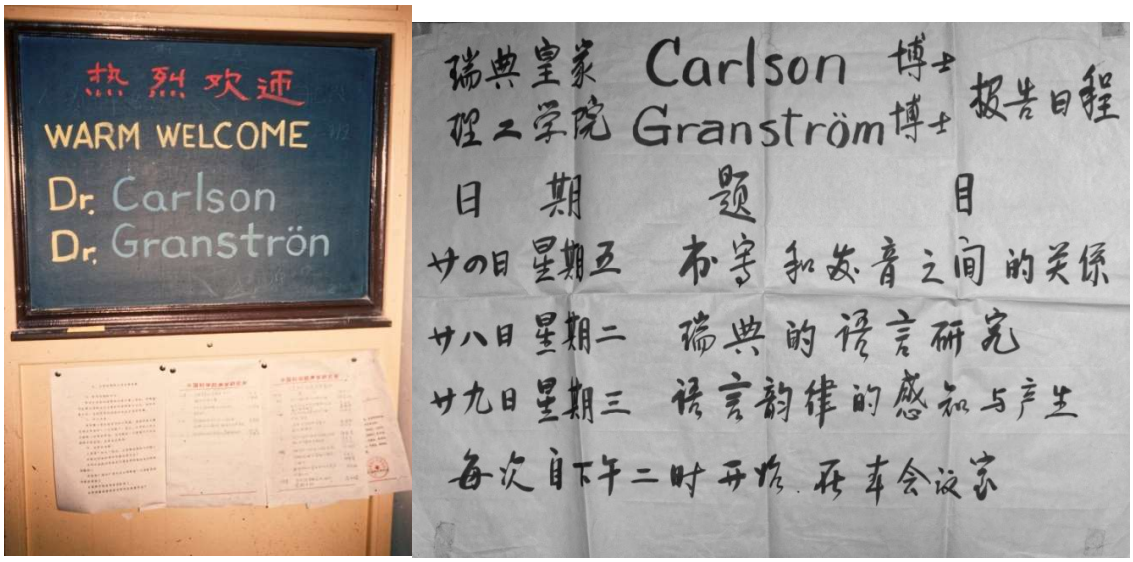
The Novosibirsk computer controlled mechanical speech synthesizer.

Details of the Novosibirsk synthesizer. The vocal tract model to the left.

Some years later, in 1980, we set off by train to Beijing by the Trans-Siberian railroad. This we made for curiosity, rather than to save the earth from too much CO2 emission. We shared the wagon with the Chinese ambassador from Austria and his wife who cooked their food on the minute stove available in each carriage. In the restaurant car the supply was running short when we approached the Chinese border. In the end the only items on the menu was rye bread (stored under the seats), Russian caviar and Russian Champagne! At the boarder the train carriages were lifted up and the bogies we changed to a wider gauge and a Chinese restaurant car was added. It was immediately completely filled by Chinese customers.

The trip was supported by the Swedish engineering academy and contained a series of lectures and study visits at several Universities in China. At that time China was quite closed and had been quite isolated during the Cultural Revolution. The young researchers did not master English, so all our presentations were translated, sentence by sentence, a very strange way of giving a speech. Sometimes it was obvious from the faces of the audience that the translations sometimes did not make sense. Then the legendary professor Wu jumped in and explained. He was educated in the US, possibly in the 40ies. Looking at the technology it was obvious that Chinese researches, during the Cultural Revolution, lost the whole mini-computer era. But now it was microcomputers, and some ASR researchers found one in a donated Bruel and Kjaer sound analyzer. And they reprogrammed the microcomputer to perform ASR!

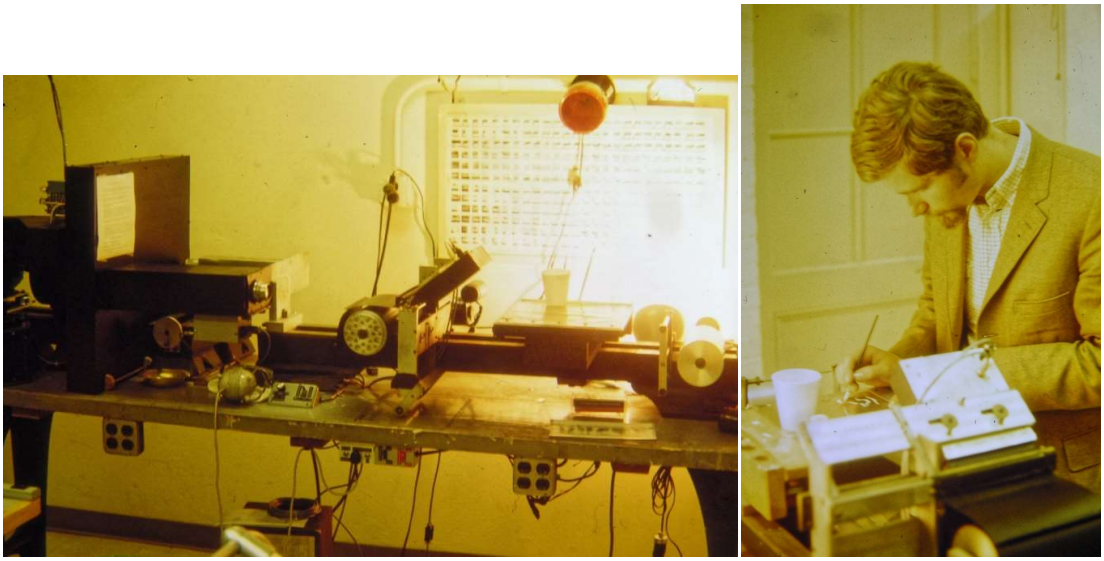
Visit to China in 1980



Visit to China in 1980

And west

In 1972 we made a trip to USA, partly together with Gunnar. It was an amazing trip. When we visited labs, young students were pushed forward to shake hands with Gunnar and many doors were opened even for us due to our company. We started to understand how big Gunnar was in the research field. And still, we shared a double room with an extra bed in a hotel in New Haven. We visited the famous Haskins laboratory and celebrated Thanksgiving at Al Lieberman's house together with Franklin Cooper. During the visit we also tried out the legendary Pattern Playback that was built by Franklin Cooper in the late 1940s. And we were put to a test: By drawing a spectrogram-like graph we were able to generate reasonable speech synthesis of our names.



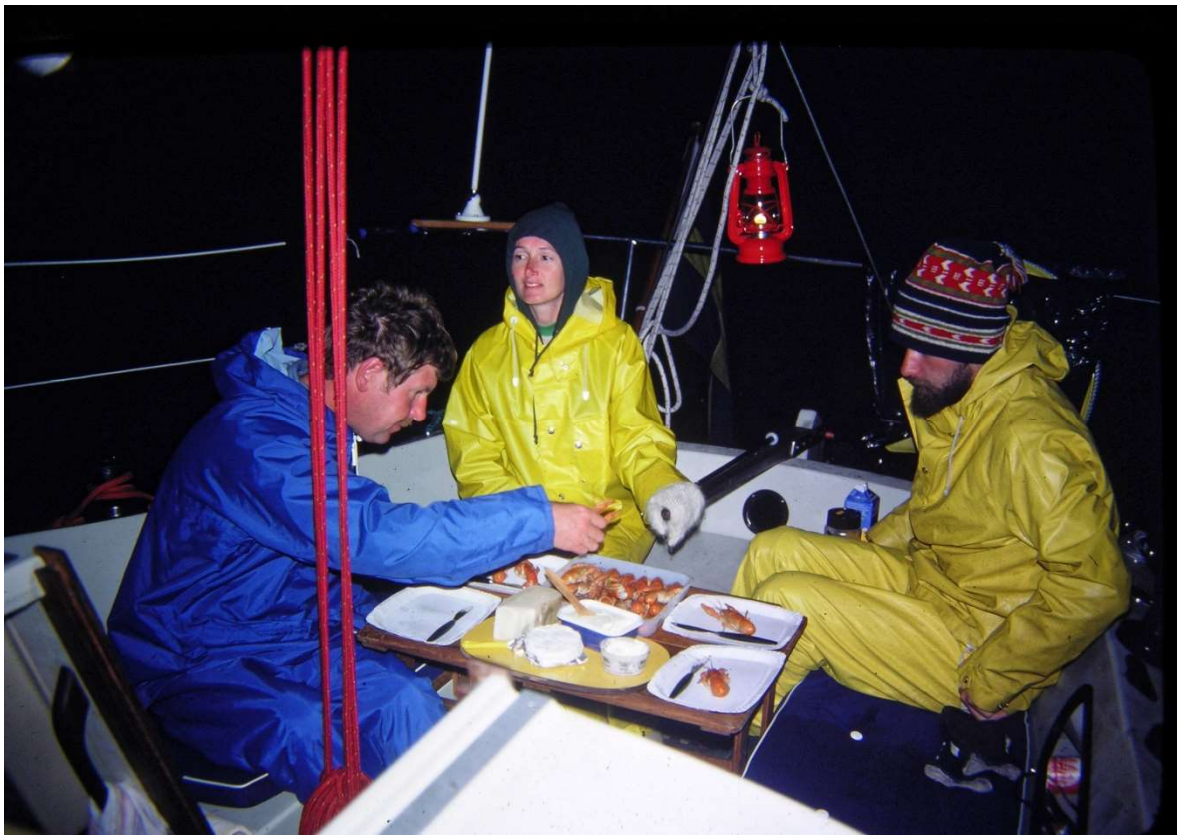
Rolf drawing his name on the Haskin´s Pattern Playback

This was also the first time we heard about the McGurk effect, maybe an early inspiration for our later research of multi-modal speech. The trip continued to Boston and MIT to visit Ken Stevens, Roman Jakobson and Morris Halle.

On water

We often arranged our travel to combine work and pleasure in an inexpensive way. We were avoiding expensive single room conference hotel accommodations and rather rented a flat or even a whole house, for the whole group from the lab. In this way more of us could afford to participate. When Björn and his family bought a 28 foot sailboat late in 1977 new possibilities emerged. Coming back to KTH after the MIT year, ICPhS, the International Congress of Phonetic Sciences was organized in Copenhagen. We, Björn, Sheri and Rolf, conveniently stayed on water centrally in Copenhagen harbor, and we convinced Gunnar to start his vowel workshop on Gotland (close to his summer house), a few days after the close of ICPhS to give us time to sail. On the first part of the journey across to the south of Sweden we were joined by MIT friends.

Sailing to Gunnar Fant´s vowel workshop on Gotland after ICPhS in Copenhagen, 1979.
From the left. Dennis Klatt, Ken Stevens, Pat Kuhl, Björn, Joe Perkell and Sara Hawkins. Except Björn all in the picture disembarked in in the south of Sweden. Björn, Bill Ainsworth, Sheri Hunnicutt and Rolf continued to Gotland.



Crayfish night party in the middle of the Baltic. The picture did not register the Nordic light

We made several trips to Finland, both to Turku and Helsinki in the same fashion. And sailing was of course a nice means of transportation for several meetings in the Stockholm archipelago.

Evening conference sail outside Turku From the left: Björn, Johan Sundberg, Anders Askenfelt and Åsa Nilsonne.

Floating synthesis show room. On the occasion of Human Factors in Telecommunication, Helsinki 1983. Sheri Hunnicutt and Paul Kiparsky

Cooperation with MIT

The researchers at M.I.T have had a special role in our international network as scientific and personal friends. At the Stockholm Speech Communication Seminar in 1974 Dennis Klatt was one of the participants. He visited us before the meeting and enjoyed playing with our speech synthesis system. At MIT, the MITalk system was under development by Jonathan Allen, Dennis Klatt and Sheri Hunnicutt. After our graduation in 1977 we were invited by Jonathan Allen to join the project and we set off to Boston for a year of exciting research. Our background fitted well to contribute to the project. We worked on all levels of the system programmed in the first modern language BCPL, a forerunner to C, which also was under development at MIT. During this time we communicated with TMH through letters and very expensive phone calls but the ARPA net was just implemented and it was amazing to see it used. It was hard to anticipate which revolution it was going start for the whole world. At the end of the year the MITalk had made good progress and was presented at conferences, an international speech synthesis school at M.I.T and in the book From text to speech; The MITalk system. The research at MIT also became the origin of Dennis Klatt's DECtalk, which Steven Hawkins used as his voice for so many years.

Rolf Carlson and Mette, Mikael, Cia and Björn Granström. Excursion to Acadia National Park, Maine, in front of our Chevrolet Malibu.

We also managed to do some parallel work on vowel perception. It was argued whether humans were able to perceive phase information. Together with Dennis Klatt, we made a follow-up experiment of our vowel perception work with experiments on phase shifted partials in vowel stimuli and proved that phase was perceptually audible under certain conditions. Also, we showed how partials in the spectrum could be deleted based on our auditory model. We stayed close friends with Dennis and despite our friendly competition to develop a TTS system for general use we kept in contact and shared our views, experiences and successes. After our year at MIT Sheri Hunnicutt came to Sweden as guest researcher and later became part of the staff.

The following years we visited MIT many times. At Ken Stevens' 60th birthday in 1984 we, together with Kjell Elenius and Mats Blomberg, presented a speech recognition experiment using an auditory model with promising result. The result was not ready when we took off from Sweden, so we did the final calculations using a statistical book in a book store at Harvard square.

Rolf and Sheri were invited by Victor Zue in 1990 for a second year at MIT Computer Lab. This time the focus was speech recognition based on auditory models and also the now emerging research on dialog system. We will discuss this exciting new research direction on human-machine interaction in a section below. In addition, during this period, Rolf and Ken Stevens spent one evening a week to discuss speech in general and to make synthesis experiments. It was precious moments of joy.

Long term visitors and short term guests

Already when we started, the department had a steady flow of guests from all over the world. In retrospect we probably met a large portion of the now “famous” persons that formed the field. It has been our ambition to follow this tradition. Long term stays by persons like Martin Rothenburg, Pierre Badin, Julia Hirschberg and many more have greatly stimulated the research at the department at different time periods. The list could be made very long and we very much appreciate of the interactions with our research friends from all over the world Despite the excellent possibility to use internet in different shapes we are strong believers in face-to-face interaction.

From acoustics to multimodal communication

During our time at TMH we have been part of a dramatic shift of research focus, from detailed descriptions of acoustic speech production and perception to a wider study of multimodal communication by man and machines. As pointed out above much of this expanded area has been possible by the rapid development of technology. Not only the rapid growth of computer capabilities both concerning hardware and software, but also other tools, like the now easily accessible motion capture techniques and MRI.

The understanding of speech communication as multimodal was obviously acknowledged also in the early days. In our hearing technology group, much of the effort was of course to compensate for the problems in the acoustic channel. Many of the speech training aids for hearing impaired persons used devices for visual or tactile feedback.

The big step towards multi-modality was around 1995 when two new doctoral students started to model the interior of the vocal tract (Olov Engwall) and the exterior “talking head” (Jonas Beskow).

One of the first projects that supported this development was the Teleface project, a proof-of-concept project that we developed with our hearing technology group. The simple idea was that a talking face could support speech perception for hard of hearing persons, in the same way as “lip reading” does in fact-to-face communication. Obviously the important visual information lies not only in the lips and contains much more than phonetic information, like attitudes, emotions and turn taking signals. It was easy to show that audiovisual synthesis was better perceived (in e.g. noise) than audio alone. The big challenge in the project was to derive the articulation or rather face movements from natural speech. When used in telephone conversation, which was the ultimate aim of the project, only a short delay could be tolerated. The project was nominated as one of the (three) most innovative projects for disabled persons. After a successful end of the project an EU project, Synface, was granted, also with industrial participation.

In parallel with this development the interest in supra segmental phenomena increased, both verbal and non-verbal, like attitudes and emotions in speech. The talking head has been used in many research projects on e.g. human machine interaction and in numerous applications including our early dialogue systems and systems for language learning. The talking head models have always been 3-D, originally displayed on 2-D screens. A quite drastic step forward was taken when

the models could be backprojected from inside facial masks within an EU robotics project. This device was commercialized and is the base of the Furhat company and is now used in many social robotics projects also in other laboratories.

The dialog system Waxholm

In science the time at some point is ripe for the next step. We have seen in many cases that suddenly the research takes a new direction. Perhaps we understand that what we know and have been doing so far now can be combined in another way or that new developments in another discipline opens a new exciting research path in our own field.

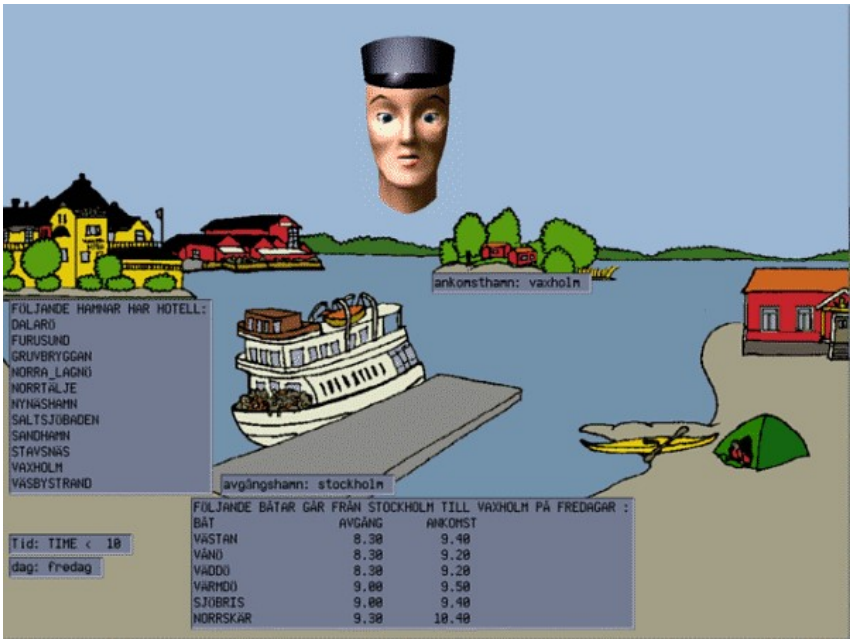
In the beginning of the 90ies such a cross fertilization happened. So far, we had mainly been concerned with basic questions such as speech perception or speech production in the field of speech recognition and multimodal speech synthesis. We had reached a breakthrough in making speaking devices available for persons in need and for some limited public applications. But it was clear that we could not just put the speech synthesis and speech recognition systems from the lab, or from the Infovox company, in the hands of developers and hope that speech technology should easily be put into public use. A completely new way of interacting with machines was needed, far from sitting at a text-based terminal. We had to create a new human-like approach based on human interaction models.

At some places in the US the first steps towards spoken dialog systems were taken in the beginning of the 90ies with support from ARPA (Advanced Research Project Agency). We were greatly inspired by the ATIS project dealing with airline ticketing and the Voyager dialog system at MIT, focused on navigation. We gathered all members of the TMH speech group and proposed a joint research effort to create a complete multimodal dialog system. The Waxholm system, dealing with boat traffic in the Stockholm archipelago, was born. We are very thankful for the visionary views of the grant giving agencies during this period. The HSFR and Vinnova had decided to jointly support a Swedish National Language Technology Program and our dialog system proposal was accepted and we received a rather generous support. This project financed a large part of the speech group, with the mission to create a functional multimodal dialog system based on human interaction.

The success and visibility of the Waxholm project and the follow up projects formed our next 15 years of research to a large extent. It led to further support such as a Telia sponsored professorship in language technology in 1995. Rolf was appointed to this chair. In 1996 the CTT competence centre was created.

Once the financing had been secured many technical details needed to be solved. A multitasking system was developed in which each module contributed output to a blackboard which was processed further by new modules using this information as input. It became easy for each researcher to focus on one aspect of the system. A special feature was that the information flow was recorded making it possible to replay a complete dialog with the modules in action for debugging and demonstration. The multimodality and the partially statistical dialog modeling were features that together with speech input/output modules made the KTH system unique and a pointer to the future.

Joakim Gustafson took part in the Waxholm project already from the start and when it was time for the next generation of dialog systems, like August and Urban, the dialog system group was successively expanded with Linda Bell, Anna Hjalmarsson, Gabriel Skantze and Jens Edlund. We now had a team in the front line of human-human interaction and dialog system research. It is so rewarding to see how the department has kept this strong position in this now very hot field of AI research.



Waxholm user interface including the multimodal agent, 1993

CTT - Centre for Speech Technology (Centrum för Talteknologi)

In 1996 Vinnova, a grant agency in the science and technology area created the competence center initiative. It aimed at supporting development in strategical areas by cooperation between universities and other organizations. The support was planned for a maximum of 10 years. This was a funding scheme previously unheard of in Swedish research, where we normally had to rely on short term grants with, at the best, three years duration.

At the onset of CTT, speech technology within Swedish industry was quite limited and was concentrated to a small number of companies, mostly within telecommunication. During the ten-year duration of CTT, this picture changed considerably, reflected by the increased number of CTT partners. Four partners during the first stage increased to 18 during the last stage. In total, 34 companies and organizations were members during the ten-year period. They formed an interesting mix of large companies, SMEs, and non-commercial organizations, representing many different application areas of speech technology. The strategy, to ask for in-kind-contributions from the partners, was very successful and increased the mobility and interaction between the partners. Several of the partners today are companies with former CTT employees. The in-kind contributions were mainly in work, but also in tools and resources important for the CTT operation.

The emphasis on CTT tools and resources resulted in the availability of both speech and language databases, some of which are multi-modal, and in an extended toolbox. These resources have successfully been used in many experimental applications and demonstrators. Especially Snack and WaveSurfer, created by Jonas Beskow and Kåre Sjölander, published as open source, received great appreciation worldwide with thousands of downloads.

In the KTH international evaluation report 2008 we could read “The VINNOVA Centre for Speech Technology (CTT) provides a platform for cooperation and project funding with industry. ….. This is an outstanding, world leading research group – among the top and most respected (a national asset).”

Graduate education

At a technical university, like KTH, graduate education is one of the most important and also pleasant responsibilities for departments. As we already mentioned in the introduction the path towards the doctor degree has drastically changed. When we started, the focus was to do high-quality and relevant research with some broadening studies. It was fully accepted to continue research and the formal examination was secondary. Today the graduate education is much more structured and a doctor title is a necessity towards an academic or industrial position. This means that we need to find support for a graduate student during a limited period of time, for mostly focused research and preferably of industrial relevance, and still give time for studies and more visionary views.

Several research projects enabled us to create such necessary environment for our graduate students. The synthesis project, the Waxholm project and other externally funded national research projects, a multitude of European projects, and finally the competence center CTT successively opened up for more and more graduate student positions.

As CTT was very successful in attracting leading scientists from Europe, USA and Japan we also engaged them to give what we called "Bullet courses" which were very much appreciated since they often presented cutting edge research. These courses also gave credit points in the student’s curricula. At the end of CTT we could sum up to 20 exams at licentiate or doctor level and also 82 MSc theses, a result the senior researchers at the department should be very proud of.

An important step in education was also the creation of the Graduate School in Language Technology (GSLT) hosted by Gothenburg University (2001). We had for some time together with the universities in Gothenburg, Uppsala and Linköping been arguing for a national platform for graduate training in language technology. The new school provided a broad interdisciplinary platform for graduate education in language technology. The unique and visionary initiative created not only higher education with joint classes but also a national meeting point. The government supported not only the school infrastructure but also the education for a number of students distributed over departments at ten universities and colleges in Sweden. It included departments of computational linguistics, computer and information science, library science, linguistics and phonetics, philosophy, speech technology and Swedish language.

During the typical five years of graduate studies it is so rewarding for a PhD thesis supervisor to follow each student's development towards a mature researcher and often also a mature human being. You follow how they successively take command of the research field, how new ideas are created and how their research network is expanding. To sit in the audience and experience how their first presentations are turned into well attended talks at international conferences give you such a great feeling and even pride. When they later graduate your feelings are in a way mixed. Do we have to lose the daily contacts? Are they going to stay in the research field? Are they actually going to stick to the academic career, or will they carry the knowledge into the industry? The rapid development of applied Speech Technology or Artificial Intelligence has made it possible for most of our graduate students to continue their work in the field, and in several cases this has created new jobs in start-up enterprises, like the Furhat company.

Final Remarks

As already mentioned in the introduction, our ambition has not been to review all aspects of the last 50 years at the Department of Speech, Music and Hearing at KTH. Rather it is a personal view of our life at this fantastic place. It has been a privilege to work with so many excellent researchers and so clever colleagues. Many are now close friends. This also goes for other colleagues in Sweden and all around the world. We have not been able to mention everyone that has been important for us. A few names have been included, but many are not.

During these 50 years we have experienced so many great moments of joy at celebrations, graduations, Christmas parties, CTT parties, retreats, riding classes, soccer matches and many other activities besides work. This is perhaps what one can expect from a functional team where the atmosphere is to support, inspire and recognize. We are proud of being part of the TMH family so well initiated by Gunnar Fant. We have, as best as we could, tried to continue this tradition and to make it a good workplace where people do excellent research and enjoy being there.

We look forward to all activities that the new generation, headed by Joakim Gustafson and Jonas Beskow, are engaged in. The group produces excellent and recognized research. Researchers are already invited speakers at prestigious meetings and take active and sometimes leading part in national and international networks. The future appears very bright and we wish the entire group a great success. Perhaps one can even say that the future already is here.