A look at 2000

Anders Askenfelt
Chairman

During 2000, a notable anniversary occurred which gives us a perspective on the activities that took place during the year that passed. In the early 60´s, several speech research laboratories in the U.S. launched scientific report series which were distributed periodically. So did the Speech Transmission Laboratory at KTH, the forerunner of the Department of Speech, Music and Hearing. In October last year, 40 annual volumes of our Quarterly Progress and Status Report (QPSR) had been published and distributed around the world, reaching more than 700 subscribers. Our report series continues, but the printed edition will be replaced by an electronic on-line edition starting next year. This anniversary reminds us that a substantial period of time with intense studies in speech communication, music acoustics, and hearing has passed. The change in distribution technology is yet another reminder that new media for text, sound and picture have become widely used. These new media use and distribute speech and music in novel ways, and naturally our research is heavily influenced by this development. Still, however, many basic questions related to the production of speech, sound generation in musical instruments and to the perception of speech and music, remain unanswered, needing continued studies.

In the early days, the base technologies speech synthesis and speech recognition were the topics of advanced research projects. The advent of digital computers made a massive approach possible. Today both speech synthesis and recognition are off-the-shelf products, available from several manufacturers, but still the systems need considerable research efforts to improve the quality. A major challenge now is to build systems in which the benefits of communication by speech are exploited, often in combination with vision or other modalities. The development of such multimodal spoken dialogue systems is a major field of research within our speech technology research group. Test cases and round-the-corner applications are of course plentiful. Searching for a new apartment is a common need in large cities (including Stockholm), and browsing apartments on the web in a conversational dialogue with an animated agent may certainly save time and effort. This type of service is currently being evaluated in a pilot project.

The complex nature of spoken communication as displayed in conversations between humans indicates that there is long way to go before a man-machine dialogue will feel natural. During the coming years many challenging tasks must be approached within this research area, but there is much to gain. Dialogues between humans are extremely efficient as many things need not to be said explicitly, and man-machine conversations would profit tremendously by such features. One example of an important component in speech which is far from fully understood and modelled in present systems is prosody, including intonation, duration structure, and visual cues. In a natural dialogue, the prosodic component is used for contrasting important parts of the message against less important parts, just to mention one function.

Many would claim that the understanding of the underlying meaning of a spoken message (requiring an interpretation of the prosody as one element) is a truly human quality. But the acoustic cues we give for these underlying meanings are probably sufficient for an analysis and interpretation by a computer, applicable in everyday man-machine interactions. In the future, the classical sentence from 1951 by the great pioneer in linguistics and speech research Roman Jakobson, ".... we speak to be heard, in order to be understood," may be applicable at all its levels even to man-machine conversations.

Returning to the early speech research, different ways of reducing the bit rate of telephone transmitted speech were studied intensively. Complex coders and decoders were designed in hardware (vocoders), which reduced the required transmission capacity to a set of slowly varying parameters of the speech signal. Surprisingly low bit rates of the order of a few thousand bits per second were reached. Present work in speech signal processing can reduce the bit rate in much more economic ways, while retaining the quality of speech. The preservation of the quality builds on parallel advancements in the perception of speech and speech sounds. Also an increase in quality compared to the band-limited signal transmitted over a telephone connection is possible, so called bandwidth extension. New media like the Internet introduce transmission errors different from those present in a conventional telephone connection, and new error correction strategies are required. Coding and compression of speech and audio on a perceptual basis and related issues are studied intensively in our speech signal processing group.

The communication of emotions by speech and music is an area which is difficult to study and which has defied many attempts. Forty years ago this research area was hardly touched. In particular the simulation of emotions in synthesised music has been notoriously difficult and seldom convincing. Our music acoustics group has now shown that it is possible to give an emotional colouring of a rule-based computer rendering of a piece which is "correctly" perceived by a vast majority of listeners. This is of course not to say that we are dealing with communication of true emotions. Transmitting the mood of the performance in a concise manner without misunderstandings must, however, be characterised as a major step forward in the understanding of music communication. In passing it is worth mentioning that the workhorse program for our music performance studies, the rule-based performance program *Director Musices*, was awarded first prize in the prestigious Bourges Music Software Competition last year.

Similar strategies as for communicating emotions in music are also explored in speech synthesis, using rule-based approaches. Mood-related facial expressions and other sources of extralinguistic information have already proven to be an important feature in multi-modal dialogue systems for improving interaction and naturalness. If, or rather, when more intriguing emotional aspects will be applied in everyday speech communication services, we will be faced with a behavioural question: Do we actually want to hear and see talking agents, robots and the like appeal to our deeper emotions, or is that a domain we would like to have reserved for communication between humans only? The interest for already existing cute robotic pets, primitive as they may be, may suggest an acceptance.

A serious hearing impairment is a profound handicap. Even the reduction in the performance of the auditory system which follows naturally with age restricts perception of the environment and is a hindrance for conversation in noise. As far back as in the very early issues of QPSR, experiments on hearing-impaired persons' ability to perceive speech and evaluations of hearing aids were reported. The work on advanced hearing aids which is today taking place in many laboratories all over the world, including our hearing technology group, promises better everyday life for a large number of handicapped people. In particular the power of present-day hearing aids to recognise and adapt their performance to varying sound environments, from a busy subway station to the quiet living room at home, represents a major step forward. In order to exploit these features, individual fitting of the aid is needed. How this should be done in an optimal way is, however, still far from clear, and contrasting strategies are under evaluation internationally.

Particular difficulties arise for hearing-impaired people in telephone conversations. In a joint effort between the speech and hearing groups an exciting attempt is made to use an animated agent for assisting hard-of-hearing people when talking over the telephone. The agent is driven by the spoken message and converts the sound to facial movements, primarily the jaw, lips, and eyes.

So much for perspective, reflections and the-year-that-has-passed activities. Follow our work during 2001 including voice experiments, phonetic studies, synthesis of speech and music, man-machine conversations and much more in our Quarterly Progress and Status Report on-line http://www.qpsr.speech.kth.se

At the end of 2000, a total of 58 researchers and research students and 7 administrators worked at the department and the associated Centre for Speech Technology (CTT). The centre, which has been operating since 1996, receives support from the Swedish National Board for Industrial and Technical Development (Nutek), 12 industry partners and KTH. The total turnover during 2000 including CTT was 35 MSEK.

During 2000, some of our researchers have left the department and continued their work in speech technology at large telecom industries like Telia, or explored their knowledge in start-up companies. The transfer of people is an important part of the interaction between universities and industry. In this way, advancements in engineering sciences, initially driven by university departments, is naturally transferred to R&D environments in industry. Particularly in a hot area like speech technology the industry demand for qualified people is high. We wish our former co-workers good luck in their new work and prepare for expansion by welcoming new research students, devoted to studies of communication between humans and machines by speech and music.

Meet us also on our home page which brings you the latest news on current projects, job opportunities, courses and faculty and staff http://www.speech.kth.se