SPEECH RESEARCH IN PERSPECTIVE

GUNNAR FANT

Dept. of Speech Communication and Music Acoustics
Royal Institute of Technology, KTH
Box 70014, Stockholm 10044, Sweden

ABSTRACT

The purpose of my talk will be to contribute to a perspective of speech research - to relate our present activities to the past and to the future. Much could be said on this topic but I will be neither complete in my presentation nor balanced in my judgements. I shall try to communicate feelings rather than facts and I will rely heavily on my own experience which spans almost half a century of speech research.

I shall not go very far back in the history. The real break through came in the decade 1945-1955 with the advent of information theory, distinctive feature theory and developments in electronics and acoustics. Bell Labs, MIT and Haskins Laboratories were the leading centers. The speech spectrograph gave an insight into a new world of Visible Speech patterns and was supplemented by speech synthesis techniques. This was a very creative pioneering period that formed much of present days concepts and a perspective of technical applications.

After this early period of analogue techniques we gradually entered the computer age. Computers were first used for calculations only but since the last 15 years almost all experimental techniques are computerized. The VLSI techniques have opened up potential markets for speech technology products. High system complexity may be combined with low cost in mass production. This is indeed promising and has evoked high expectations about a coming age of information technology were we may speak to computers as freely as with humans.

However, optimistic industrial forecasts have not been fulfilled. The marketing goals have again and again been readjusted in time. One could speak of a crisis not only in fullfillment of marketing prospects but also in the actual progress of technology. Have we overestimated the need of speech technology products? Perhaps, but the basic problem lies in perception. It is tempting to promise more than can be achieved within a limited contract period. This is true of individual research groups as well as national programs. Text-to-speech may gain a big market once it can compete with direct speech coding in quality. The present status of the art is remarkable in view of the complexity of the task which is promising, but we need improved performance to stimulate the marketing. There are demands of greater naturalness and a flexibility to adapt to the needs of various reading styles and speaker representation, i.e. various categories of male, female and childrens voices. Speech recognition is still in a rather primitive stage. Large vocabulary, speaker independent systems cabable of handling connected speech is still far ahead. We also need to supply both speech recognition and text-to-speech systems with extended language competence including semantic differentiations. The computer has to understand in order to react properly on a voice command and it needs some understanding to read a text properly.

Is this really within reach? Have we not underestimated the formidable task of making machines perform as humans? The answer is both yes and no. There are obvious limits but we have still long to go in order to fully exhaust the potentialities. Art present we are in a state of crisis. There will surely be a way out but there are two competing philosophies or strategies to consider. One is the knowledge approach. Our models of human speech and of speech communication processes are in many respects incomplete and primitive. We may have acquired a reasonable overall qualitative view of the nature of speech but we have not been able to organize our insights into quantitative, operative representations. We are plagued with a variability of realizations. We have a superficial insight in the overall structure of variabilities but we have not been able to formulate consistent rules within a wide frame of contexts. We are thus at a loss when attempting to specify invariance criteria. Of course, we may as in distinctive feature theory, resort to concepts of relational invariance but for practical use in recognition we anyhow have to cope with the complete conditioning frame.

Thus, with a more profound knowledge we could break the "speech code" in all its complex dependencies and pave the way for a new generation of applications.

EUROSPÉECH ’89, Paris, France, September 1989
Many research groups, including our group at the KTH in Stockholm share this philosophy. We favour a knowledge based approach. The computer might do the job but can not tell us how. There is no system to reveal, no invariance criteria, only a complex pattern of joint probabilities set by a large number of weighting factors within a complex layers of neuronlike units trained for the specific task.

In a sense we have become the victims of computers. Computers inhibit our creativity. Much time is lost in programming and debugging. We do not get the immediate feedback and interaction needed to match our pace of thinking. At present, the testing and development of models takes too long time. One day we may have these resources so that an experiment may take a day in stead of a year to complete. Inspite of the fantastic growth of computer technology we may only have seen the beginning of a development. I feel that we are still in an early stage.

Let us turn to the research issues. What are our main problems and how should we direct our efforts? First of all let us make it clear that it is not a matter of concentrating on one narrow area. We need a continued broad support for work within all disciplines related to the speech communication chain. Today, advanced knowledge based development of speech recognition is structured so as to integrate most of what we know from the linguistic level, from speech production theory, auditory oriented signal analysis, etc. This is our ambition at the KTH. However, such multilevel system architecture will not perform better we have the knowledge included. Our present models of speech dynamics are incomplete and not sufficiently free from the segmental frame.

Strategically, we would benefit much from systematic articulatory modelling. An approach in this direction is to make maximal use of known articulatory constraints in programming formant synthesis. This is more or less established practice. However, only by a complete articulatory modelling can we gain the true benefits in synthesis as well as in general understanding of the speech code, e.g. in articulatory interpretation of the speech wave. Articulatory modelling has the potential of preserving continuities and it is obviously the natural basis for a description or coarticulation and reductions. A related dimension is distinct "hyperspeech" versus reduced "hypospeech" and the general notion of articulatory contrast.

Prosodic features need to be defined on an autosegmental level free from the usual phonological constraints. In articulatory modelling prosodic and inherent features are optimally combined. Recent work in our laboratory supports the view that pauses in speech are influenced by rhythmical considerations in addition to grammatical and phonological constraints. Articulatory modelling also allows a flexible choice of speaking style and speaker characteristics. Simple resetttings of articulatory constraints and basic dynamic parameters would preserve most of the small details that we associate with a specific voice or speaking style. However, there remains much work to be done on articulatory modelling principles and to collect supporting physiological data. Such direct observations can be supplemented by analysis-by-synthesis techniques to reveal trajectories of articulatory gestures that match a spectrogram. It shall be interesting to follow the developments within this area.

The basic theme of my presentation has been the importance and the need for basic research. Although I have stressed the articulatory basis of descriptions there remains much to be done in speech perception research as a complement to the overall code. As already mentioned, we also need a more fargoing integration of phonetics with grammar and semantics. In the early part of my paper I referred to the American dominance in speech research 40 years ago. During the last 10 years we have experienced an impressive growth of European speech research, both technology oriented work and activities in the broader field of fundamental research. In the latter respect the joint European activity is not behind that of the USA and Japan.

There is an impressive number of interesting papers that have been contributed to our conference. This fact supports my views of the growing joint strength of European speech research. It also guarantees that my very best wishes for a successful meeting will be fulfillled.