Dept. for Speech, Music and Hearing

Quarterly Progress and Status Report

Speech research in perspective

Fant, G.

journal: STL-QPSR
volume: 30
number: 4
year: 1989
pages: 001-007

KTH Computer Science and Communication

http://www.speech.kth.se/qpsr

SPEECH RESEARCH IN PERSPECTIVE

Gunnar Fant

Abstract

This is an invited lecture given at the opening of EUROSPEECH 89, the second meeting of the European Speech Communication Association, Paris, September 26-28, 1989.

A review of personal impressions from 45 years' engagement in speech research leads up to the present challenging developments in speech technology. How shall we proceed? What are the benefits of a knowledge approach versus a statistical approach? Much fundamental knowledge is needed to force the speech code. Computer technology alone cannot pave the way. Models are necessary but also deceptive. An optimal interface of models and true speech data is needed in all development work. The interdisciplinary nature of speech research and its importance for a wide area of human knowledge and applications is stressed.

INTRODUCTION

The purpose of my talk is to contribute to a perspective of speech research - to relate our present activities to the past and to the future. Much could be said about the developments of speech research and speech technology. I will not be able to give a complete overview of this fascinating field, and I intend to convey personal impressions rather than facts. I will rely heavily on my own experience which spans almost half a century of speech research.

I shall have something to say about the great expectations we have to live up to, the great challenge of making computers speak and understand, if not as human beings, at least at an advanced level of performance. But what strategy shall we choose? Can we pursue a knowledge-based approach and eventually force the speech code, or shall we attempt to train our computers to learn the task by statistical inference? Shall we leave it to the computers to handle a problem we have failed to formulate and structure in a working code?

We live in the computer age in the information society. Computers have given us a tremendous processing capacity, but our symbiosis with computers is not without problems. They are our partners for the better or the worse. Also, they are no better than the models we use. Models are necessary. Models can be intriguing but models can also be deceptive, which we sometimes find when we confront models with the real world.

As you may understand I have a strong personal bias for a knowledge-based approach. With better models on all levels of the speech communication chain, including language theory, speech production, and speech perception, we should do a better job in speech synthesis and speech recognition.

As we learn more and develop a more advanced insight in the speech code, we will also be able to contribute to many fields peripheral to speech technology. Close at hand is a variety of handicap aids, but the main impact will be the contributions that we can give to overall knowledge of human functions, for its own sake, and for applications in rehabilitation and medical care of voice and hearing disorders, training of reading and writing, and, not the least, for improved methods of foreign language teaching.

This, in short, is my message. Now follow some personal notes from the history of speech research.

45 YEARS OF SPEECH RESEARCH

My own engagement in the field dates back to the fall of 1944 when I started my thesis work for the electrical engineering degree at the KTH in Stockholm. I made a theoretical and practical study of the loss of intelligibility when removing a part of the audio band in telephony to be used instead for tone signalling. This promoted my first employment at the Ericsson Company in 1945, where I was set to investigate the spectral properties of Swedish speech sounds. I constructed my own speech analysis instrumentation, a kind of wave analyzer.

I had subjects sustaining vowels for 5 seconds during which time I manually manoeuvred the sweep frequency to cover a 4 kHz range of narrow-band harmonic analysis. The output was recorded on the photographic paper of a magnetic coil oscillograph. Technically this was a great advance over what was earlier reported in the literature, where it was reported that the subject had to sustain the sound for 2 minutes or more, which demanded exceptionally trained singers as subjects. Anyhow, I got quite useful data, unique since data on absolute sound pressure levels were preserved.

For analysis of connected speech, I used the same wave analyzer with a broader filter and produced oscillographic records of the output in various frequency bands. It was a time-consuming job to organize all these data. I had just read about the sound spectrograph at Bell Labs and now I attempted to produce stylized spectrograms of Swedish stop-vowel-stop syllables, as you can see in Fig. 1.

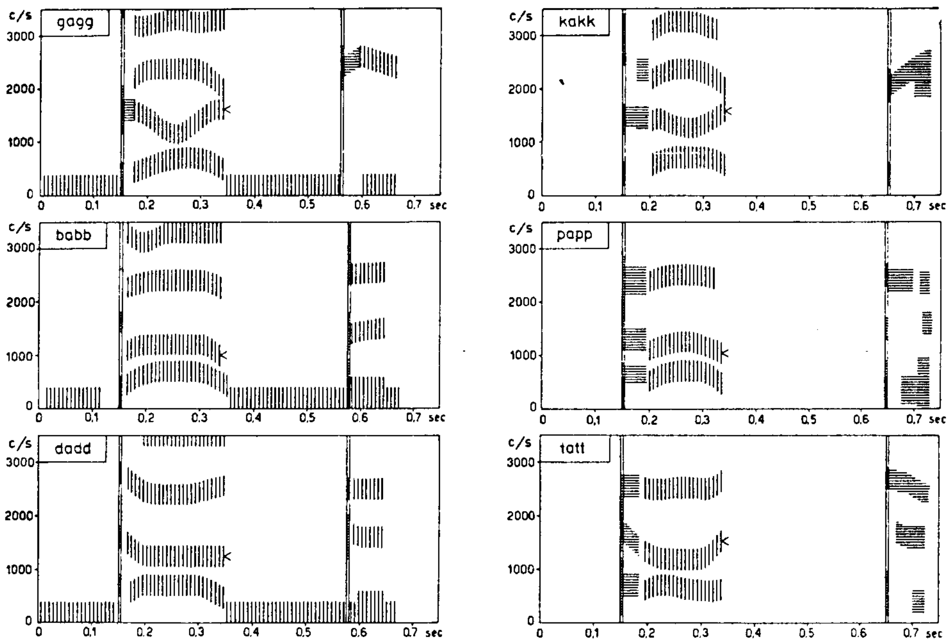

Fig. 1. *Stylized spectrograms of "gagg, babb, dadd, kakk, papp, tatt" compiled from band-pass oscillograms. The experimental work was undertaken in 1948. From Ericsson Techniques, no. 1, 1959.*

They retain the well-known cues of dentals, velars, and labials. I also produced intensity-frequency cross sections. In Fig. 2 is an example of the velar, dental, and labial unvoiced stops in syllable final position. The repeated play back was accomplished with a first generation magnetic tape recorder. In 1947-1948, I shared my time between the KTH and Ericsson. One of my projects at the KTH was to investigate the relative frequency of occurrence of phonemes and words in telephone conversations. The Swedish Telecom lab produced the recordings on an old-fashion wire recorder. This was the antique technique prior to magnetic tape recording.

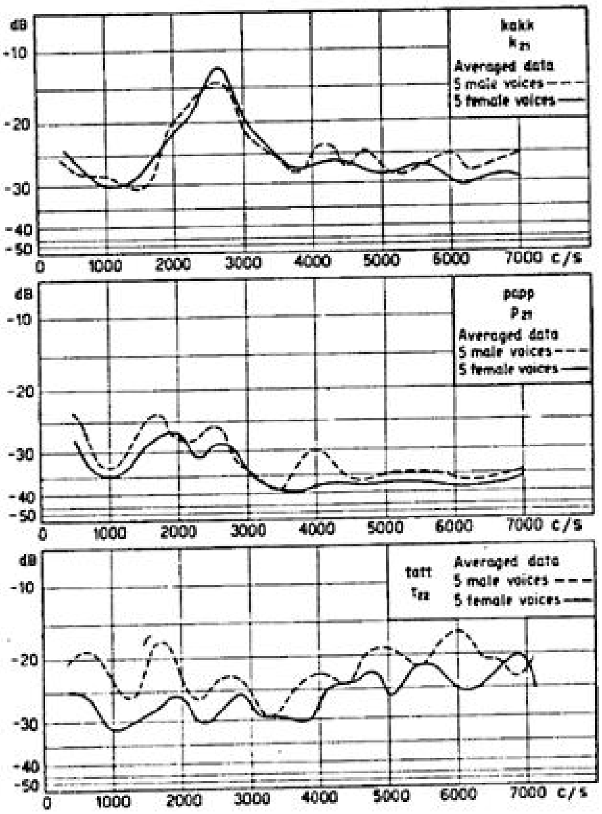

Fig. 2. *Spectral sections of the release phase of syllable final unvoiced Swedish stops compiled from band-pass oscillograms. The experimental work was undertaken in 1948. From Ericsson Techniques, No. 1, 1959.*

Anyhow, when I came to the USA in 1949 for a two-year-stay, I was well prepared to join a very active and creative period in the history of speech research. I worked for a couple of months at the Psychoacoustics lab of Harvard University and then came over to the MIT Acoustics lab. People were enthusiastic about information theory and its connection to speech communication models.

Haskins Laboratories' spectrogram Play Back synthesizer was a fascinating experience. With my accumulated knowledge from speech analysis I could directly go ahead and paint a stylized pattern of a word or a simple phrase and have it replayed. This historical age of analogue techniques provided resources for interactive experimentation at a level that we are now only slowly beginning to reach. At MIT I met Ken Stevens and James Flanagan, and the linguists Roman Jakobson and Morris Halle – the start of a life-time friendship and a growth of mutual interests. Roman Jakobson had a pronounced influence on my scientific interests and carrier. Incidentally, the spectra of Swedish stops, shown in Fig. 2, intrigued him very much as an example of the realization of distinctive features. With Roman Jakobson's help I started my experiments on X-ray photography of a Russian exile actor. His articulations are still an often used reference in speech acoustics.

The Bell Laboratory work on spectrographic speech analysis performed by Potter, Kopp, Green, and Steinberg and by Gordon Peterson was another source of inspiration.

Back in Sweden in 1951 I expanded on my speech acoustics work and formed a research group, which was referred to as the Speech Transmission Laboratory. The next active period in the history of speech research was in the late fifties and the early sixties. From this period, we may note the parametric speech synthesizer of Walter Lawrence in England followed by our OVE II in Stockholm and the pioneering synthesis work of John Holmes. He was with us in Stockholm in 1961. His and our Hi Fi synthesis experiments demonstrated that a high quality synthetic replica could be made of a natural utterance, thus establishing the potentialities of formant synthesis. The next bench-mark in synthesis was in 1975 when my colleagues Rolf Carlson and Björn Granström demonstrated a multi-level integrated programming language for text-to-speech conversion. The work of Dennis Klatt and others at MIT developed at the same time. In this connection I want to remind you of the great impact Dennis Klatt's work has had in several branches of speech research. In the last few years, there has developed an interest in multi-language synthesis, a speciality of text-to-speech development work in Sweden.

THE CHALLENGE OF SPEECH TECHNOLOGY

Back in 1970 the computer age was already well established and analogue techniques were substituted by digital techniques. In the last 15 years VLSI techniques have opened up potential markets for speech synthesis products. High system complexity may be combined with low cost in mass production. This is indeed promising and has evoked high expectations about the future of speech technology. Here is an old friend, Mr Speech Technology, (see Fig. 3) who is boldly sailing towards the wonderful island of informatics where we may speak to computers as freely as with human beings. But he is not aware of the reefs of the knowledge barrier where he might get stranded.

Fig. 3. Mr Speech technology and the knowledge barriers.

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 the fulfilment 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 performance. 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 children's voices. Speech recognition is still in a rather primitive stage. Large vocabulary and speaker-independent systems capable of handling connected speech are 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.

THE KNOWLEDGE APPROACH VERSUS THE STATISTICAL APPROACH

Is this really within reach? Have we not underestimated the formidable task of making machines perform as human beings? The answer is both yes and no. There are obvious limits, but we still have long to go in order to fully exhaust the potentialities. At 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 have anyhow 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. Many research groups, including our group at the KTH in Stockholm, share this philosophy. We favour a knowledge-based approach. The competing strategy, now adopted by most speech technology groups is statistical. Hidden Markov processes have become an established technique. The new intriguing approach of developing "neural nets" that are trained to perform a recognition task has attracted a lot of attention not only in the speech field but also in other applications of pattern recognition. It may prove fruitful, but at the same time it implies a failure. We leave it to the computer to learn what we have failed to understand. The computer might do the job but can it 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 complex layers of neuron-like units trained for the specific task.

THE FALLACY OF COMPUTERS AND MODELS

Computers have become an integral part of our daily life, our indispensable resources. But they demand so much of us - not only in large investments of capital, but also the large time spent in programming and debugging. In spite of the rapid developments I find that we have not still reached an acceptable level of performance in the simulation of complex systems. It takes too long time to transfer our knowledge into an operative form. We do not get the immediate feedback and the degree of interaction needed to match our thinking. I feel we are still in an early stage of development. We should not have to wait for months or a year to test a new idea.

Our children are brought up with TV and computer games, and we are sometimes concerned about their loss of contact with real life. For us adults in the speech lab, the situation is similar. Perhaps we should spend less time in front of our computer screens and devote more time to creative thinking and discussions with our colleagues? A program or a model can become a purpose in itself.

A loss of contact with reality is a danger for us all. A classical situation is often encountered in a synthesis lab. You call on a colleague to listen to your latest products and you ask him: "Is this not a fine nasal consonant /n/?" and his immediate reaction is: "I hear an /l/." Similar effects appear in recognition experiments. A system may operate well for one speaker but not for another. Primitive and constrained models are inhibiting. Whenever possible, models should be confronted with reality. The best example we have is in spectrogram reading seminars, which I find both revealing and rewarding. Revealing because I am constantly reminded of the shortcoming of my models, and rewarding because I get a chance to learn and revise my models.

Our models of sound patterns are often synthesis models rather than true speech models. However, an advanced synthesis model can be of some help in a recognition system, the real world is confronted with a model. Conversely, our text-to-speech development system in Stockholm permits a spectrum matching between a synthetic syllable and a preselected natural syllable. Here the model is confronted with the real world.

IN SEARCH OF THE SPEECH CODE

Let us turn to specific 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, from auditory oriented signal analysis etc. This is our ambition at the KTH. However, such multi-level system architecture will not perform better than 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 of 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 voice- and speaker-type characteristics. A selected choice of articulatory constraints and of dynamic programs would preserve most of the features that we associate with a specific voice or speaking style. However, there remains much work to be done on articulatory modelling and voice source dynamics 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.

SPEECH - AN INTERDISCIPLINARY SCIENCE

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 far-going 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, not the least in France, 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.

Finally, I would like to stress the close ties between speech technology and a number of related subjects, especially linguistics, phonetics, physiology, psychology, and a number of applied areas in rehabilitation and language training. Speech synthesis is much used in communication systems for the blind and as speech prostheses for speech handicapped persons. An important application is that of promoting new pedagogical techniques in foreign language training. So here is our flower of interdisciplinary sciences in the speech field (Fig. 4). We share the joy and responsibility of contributing to this joint effort of research into human speech, the most exciting of all sciences.

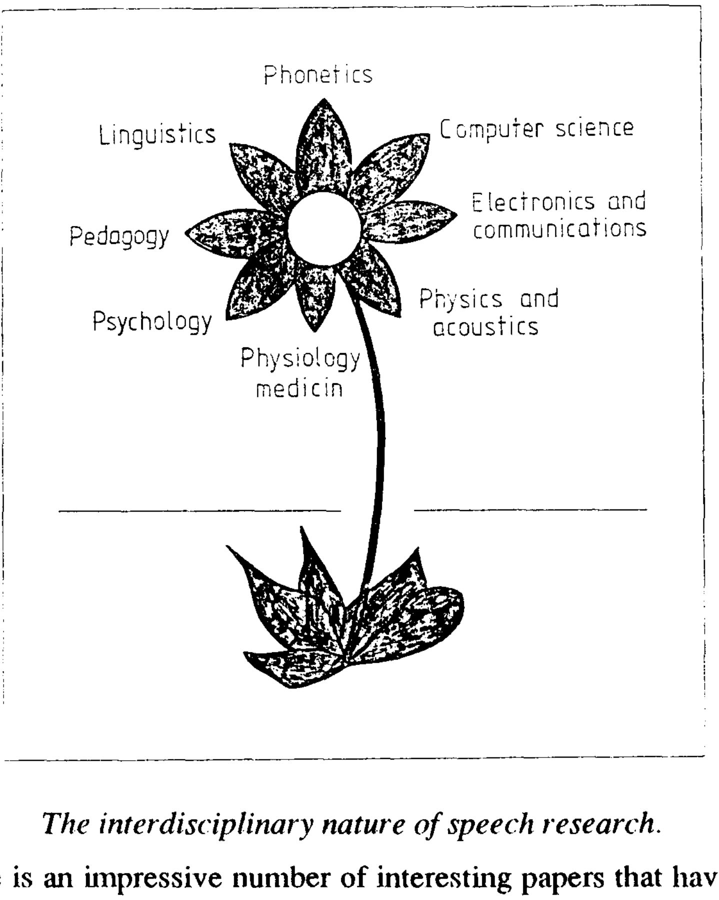

Fig. 4. The interdisciplinary nature of speech research.

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 fulfilled.