Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

The role of speech research in the advance of speech technology

Fant, G.

journal: STL-QPSR
volume: 31
number: 4
year: 1990
pages: 001-006

KTH Computer Science and Communication

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

THE ROLE OF SPEECH RESEARCH IN THE ADVANCE OF SPEECH TECHNOLOGY*

Gunnar Fant

Abstract

This is a review article on the general perspective of speech research as a fundament of speech technology. It centers around the concept of the speech code, substantiated as rules of text-to-speech conversion. Some analogies to the genetic code are pointed out. A future development towards advanced goals must in the first place rely on a profound understanding of the entire speech communication process and speech production in specific. the knowledge approach may be supported by a statistical approach to ensure maximum efficiency. The concept of contextual variability versus relational invariance is discussed. A major difficulty is to cope with a sufficient wide range of contextual rules. Contextual variabilities are exemplified by spectrographic pictures of voiced stops.

1. INTRODUCTION. VISIONS AND CONSTRAINTS

The purpose of this paper is to contribute to the perspective of speech research as a fundament of speech technology. My overview of this large and fascinating field will be far from complete. I will convey my general impressions and review some of my own work.

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

I would like to start out with a broadened perspective of our situation. Science is like a biological system, and the growth of science shows similar traits. In nature, the species develop generation after generation towards a more perfect form. The same is true of the development of speech communication systems. We are eager to build into them human functions of speech and hearing and understanding. However, in many respects the progress has not been as fast as we have hoped. Although speech technology is striving ahead with an increasing effort of people and funding, and with hard work and not without partial success, this is still along a path wrecked with technical failures and frustrations. We seldom reach the level of performance we are aiming at, and everything takes longer time to complete than anticipated.

In this situation we may need a reminder. We are trying to model and reconstruct something that it has taken nature some more than 50 000 years to develop and with a natural resource of a brain, in comparison to which our present computers are most trivial and simple tools. No wonder that our concepts of human speech communication remain primitive. Computers are indispensable but our symbiosis with computers is not without problems. They are our partners for better or for worse. Also they are no better than the models that we employ. Models are necessary. Models can be intriguing, but models can also be deceptive, something that we often find when we confront models with the real world.

---

*This paper was presented at a Workshop on Speech Technology for Man-Machine Interaction, Dec. 10-12, 1990, TIFR, Bombay, India)

Maybe, in spite of the tremendous advance of computer technology, we are still in a stage of transfer from its infancy to the middle age, at least in a perspective set 50 years ahead from now and looking back at our days' primitive models of brain functions. In spite of attempts to introduce flexible parallel processing systems and interactive operating strategies, we are still unable to operate at a pace that matches our scientific curiosity. What should take half a day, may take weeks or months or years of laborious programming work to complete. These long delays inhibit a creative thinking. We have a moral right to be impatient.

So – who is the master and who is the slave – the man or the computer? When will we be free to release the creative energy of our scientific curiosity? We are looking ahead, but in the present situation we have to carry on with limited resources and a limited insight, year after year, constantly updating our plans.

What is it that keeps us going? What are we aiming at? In a biological system, it is the genes that carry the message and the mutations that ensure a development, generation after generation in a competition where the fittest will survive. In speech research and speech technology, the situation is similar. The driving force is the impact of our scientific concepts, visions and dreams. Without visions, no advance. But visions should be based on knowledge and experience. To make our dreams come true, we must be able to substantiate them. Although somewhat remote, one could conceive of an analogy between the genetic code and the speech code. Our concepts of language and communication need to be analyzed and structured and finally expressed in a more or less discrete shape of a code suited for reproduction, for instance that of synthesis-by-rule. The future of our field depends on how well we can pursue this task of penetrating the speech code.

Our visions and ambitions vary over a wide range. The most advanced is to be found in the Japanese challenge, the dream of the interpreting telephone service that some day hopefully will enable an English-speaking person or a member of any major language group, to call somebody in Japan via a fancy translation system that not only recognizes the words and understands the message but also is capable of conveying some personal speaker and speaking characteristics, including specific denotations and attitudes to be preserved in an automatic translation to spoken Japanese.

This vision, a science fiction that may be reality in fifty years from now, illustrates four basic aspects of the speech code. As listeners we may ask ourselves:

"What was said?"
"How was it said?"
"By whom was it said?" and
"What was meant?"

Most of our efforts are concerned with correct synthesis and recognition of words and sentences, i.e., "what was said", and much less work on "how was it said" and "by whom was it said", these two aspects being referred to as extra-linguistic information. The lack of knowledge is even greater at the final level, "what was meant". Speech understanding in man-machine interface is limited to selected tasks with a constrained language. Automatic language translation for telephony is an even more difficult undertaking. In spite of the very large research effort put into automatic language translation, we are held up by the semantic barrier.

2. IN QUEST OF THE SPEECH CODE

Let us have a closer view of the concept of the speech code. There is a backbone in linguistic theory and in the structure of a particular language. We are concerned with the encoding of speech messages through the various stages of the speech chain, in production, in the speech wave, and in perception, and we finally end up at the semantic, cognitive level. A major limitation in present days' text-to-speech conversion is that the computer does not understand what it is going to say.

Phonetics and speech analysis are directly involved in the study of the code in its two major parts, segmentals and prosody with their high degree of interaction and variability. Our present knowledge is qualitative rather than quantitative and scattered rather than integrated in an overall

structure, and we are constrained by primitive models. The speech code is not only a matter of social conventions and individual habits. Spontaneous speech and dialogue have their specific codes which in part are different from text reading. Furthermore, the properties of words produced in isolation or in short sentences, that is what I call "lab speech", are not necessarily representative of connected speech.

In speech research we tend to choose narrow problem areas that suit our background and competence. This is unavoidable but may lead to fragmentation and difficulties in relating the data quantitatively to other aspects of the code. Thus, tracings of articulatory movements of pellets and other sensors attached to the tongue or the lips can provide a limited insight about an articulatory gesture. However, such data are not directly translatable to the dynamics of the vocal tract system function. An insight into the proper geometrical dimensions of the vocal tract is generally not available. Vocalic and transitional elements that have a direct association to a formant pattern have been given much more attention than complete production models that include specific consonant characteristics.

We have a diversity of interests and specialities. Some of us are more interested in defending a philosophical issue than contributing knowledge about the code. Data from narrow range studies are used as evidence for far-reaching conclusions. However, we certainly need new perspectives to stimulate alternative approaches.

A fundamental problem is to bring together the knowledge that does exist. I have spoken of a knowledge gap and a semantic barrier. But we are also faced with a cognitive gap in bringing order and structure to data. There is a limit to the amount of information we can handle in theory, just as there are practical limitations in the complexity of a text-to-speech system. There is an unavoidable pay off between capacity and simplicity. A complicated model or theory is difficult to grasp and to realize, whereas simplifications may lead to information loss. We have to rely on an optimum degree of complication in our modelling, and we have to rely on a maximally efficient structuring of the speech code.

We cannot develop the code along a concept of absolute invariance, e.g., for recognition purposes. It should be remembered that the distinctive feature theory of Roman Jakobson is really a theory of relational invariance. Irrespective of context, a minimal contrast in one feature retains some aspect of a common denominator but the phonetic nature of the contrast is, of course, dependent on the particular context. Statements concerning a common denominator should retain sufficient phonetic reality to serve as a distinctive feature correlate within a phonological system. However, the common denominator easily becomes diluted and unspecific and, thus, insufficient for recognition purposes. Instead, we should concentrate our efforts on deriving rules for contextual variabilities within a proper linguistic framework. The concept of distinctive feature as a class of phonetic categories and their relational contrast is helpful in this respect. We should also consider the two aspects of the attribute "relational". One is with respect to the immediate context of preceding and following phonetic events of the speech wave. The other is in relation to the ensemble of possible alternative features or phonemes that could occur in the particular context of already recognized or alternative language units. This is the difference between dynamically varying cues and top-down biased selections. Both aspects belong to the speech code.

A major source of difficulty in structuring the speech code is the high degree of interaction between simultaneous conditioning factors. Our knowledge is incomplete, and we lack a comprehensive framework. How shall we cope with this situation? Most people would back up my proposition that we should develop the code along a speech production model. Articulatory modelling and a more sophisticated use of perceptual constraints will pave the way. The segment-by-segment structure that we may observe in speech is incidental and often inadequate for phonemic segmentation. A reference to simultaneous and partially synchronized articulatory and phonatory gestures is more fundamental. Thus, instead of worrying about the relation of each of a large number of terminal synthesis parameters or speech wave descriptors to each possible message unit, we may benefit from an intermediate level of production modelling, allowing a data reduction retaining code essentials and, thus, more compact and manageable specifications.

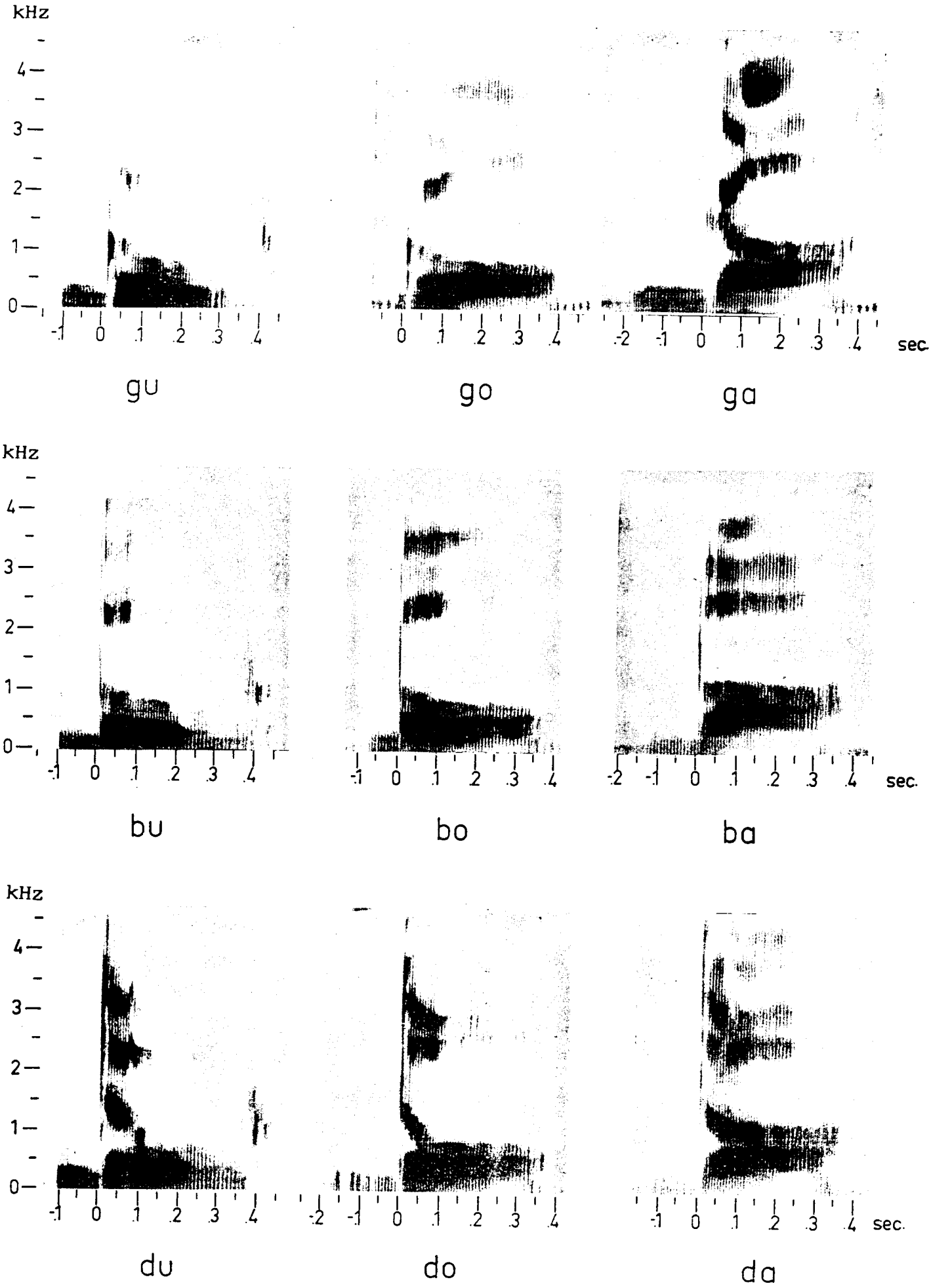

Fig. 1. Voiced stops /g/, /b/, /d/ in back vowel contexts, /u/, /o/, and /a/. Swedish informant.

3. THE KNOWLEDGE APPROACH VERSUS THE STATISTICAL APPROACH

I would like to say a few words about practical applications and, in specific, something about a knowledge approach versus a statistical approach. The relative success of present days' text-to-speech synthesis could create the illusion that we already have a most advanced insight into the speech code. However, in listening to speech synthesis, we have the obvious support of a top-down prediction which conceals apparent shortcomings. In speech recognition, on the other hand, the demands on a working code are much greater. We have as yet not been able to establish a quantitative descriptive system of phonetic pattern recognition which would substantiate an ideal expert's knowledge. As already pointed out, our knowledge is incomplete and largely qualitative. How do we cope with this situation?

The challenge comes from the statistical approach. Today, most development work on speech recognition is based on statistical methods of signal processing. Hidden Markov processing and neural net simulations have been introduced as shortcuts to the speech code. I am not very well oriented in these techniques, but they seem to involve a brute force approach. It seems to be a vaste of data and not very efficient to base a recognition on how well each single frame fits a possible message. From information theory, we know that unstructured redundancy acts as noise.

The reason why the alternative, the knowledge approach up till now, has not managed to provide the same recognition scores is, as I have pointed out, the lack of a rational, that is, a lack of concrete, sufficiently accurate and complete acoustic-phonetic descriptions of a specific language. The knowledge is, in part, hidden in the trained memory of our computers but computers cannot reveal the code. However, I am not against the use of probabilities, and I feel that neural networks have important applications for specific tasks.

The situation is similar to what I have just discussed in connection with distinctive features. Given the assumption of a certain context of preceding and following phonetic units and the whole arsenal of conditioning factors, there is a finite number of alternatives within an ensemble from which to choose. In a binary choice situation, the relation of the correct choice to the alternative defines the acoustic correlates of a distinctive feature in that context. If the choice is between alternative phonemes or syllables or morphs, the number of alternatives may be limited by top-down constraints. The identification will be facilitated by breaking down the pattern into smaller phonetic units.

If we employ neural networks for this purpose, they must operate on a time-pattern matrix to enable a running analysis of phonetic events, and they must be trained to operate under a variety of contextual conditions with specific bias for specific assumptions concerning the context. The ideal network should be able to make predictions both of the particular message unit and the context. But here it is obviously an advantage to combine knowledge of the code and its implications with automatic detection. We cannot rely on neural networks alone to do the job. We must share our knowledge with them. These are just some general reflexions. Of course, one should not be limited to structures and probabilities within a short-time context. Parallel processing, iteratively within a whole sentence on many simultaneous levels, is needed. Also, the identification should not be limited to segment-by-segment determinations but more effectively to the identification of sequences of phonetic events with reference to underlying articulatory patterns.

4. A SPECTROGRAM CASE STUDY

How can we develop a knowledge basis? A main source of insight is through careful studies of spectrograms and associated speech parameter displays. In Fig. 1, I have chosen to illustrate typical patterns of the Swedish voiced stops /g/, /b/, /d/ in a back vowel /u/, /o/, /a/ context. In speech synthesis we often encounter confusions such that /bu/ is heard as /gu,/ and /bo/ is heard as /go/. Indeed, the spectrographic patterns reveal similarities that may explain these confusions. The F2 initial values and main trajectories within the /u/ and /o/ vowels are almost the same after the velar and labial stops. However, these transitions mainly reflect the progressively increasing liprounding and the main tongue body shift within the vowel. The distinctive difference lies in the release phase which, in case of the velar, shows a concentrated burst at the initial F2. In the labial there is no

burst, and we see the typical 2 ms brief release transient which is diffusely spread in the region between F2 and F3. The compactness feature of the velar is realized by the strong concentration of burst energy in the F2 region, whilst the labial release is initiated by a diffuse transient only.

A frequent mistake in synthesis is to shape the labial burst with low frequency dominance and to give it too high intensity. Both these characteristics will promote a velar response. The spectrographic patterns of the stops in the /a/ vowel context show these well known and distinct prototype patterns of the high F2, the low initial F3 locations at the velar release, and a neutral F2 and F3 for the labial. The distinctive high frequency dominated burst of the dental /d/ is also apparent. In less flexible synthesis schemes, the velar /g/ in /gu/ is given the same relative high initial F2 as in /ga/ which, accordingly, conditions a /du/ response.

This example illustrates the danger of relying on prototype patterns that are not modified by contextual rules. The distinctive elements of a sound pattern are often confined to very brief intervals and are often lost in a conventional frame-by-frame analysis. Overall "vocalic" transitions may be confused with parts carrying consonantal information. Similar intricacies occur in the place features of nasal consonants. A reference to a proper production model is a valuable guide for correct handling of contextual variations.

Those who are interested in pursuing a knowledge approach should invest much time in spectrogram-reading seminars. Personally, I find them extremely useful. I am constantly being reminded of the shortcomings of my internalized models and my ignorance of the great variability. Yet, there is a system to be learned. The spectrogram reading courses of Victor Zue at the Massachusetts Institute of Technology is a good example but in my belief, one can pursue this pedagogical mission at an even more advanced level to bring out more of the structure of the speech code and contextual variations.

5. FINAL REMARKS

This overview on the role of speech research in advance of speech technology focusing on the need of a solid knowledge base, the "speech code" for synthesis and recognition, recapitulates a theme that I have recently treated elsewhere, Fant (1990a; 1990b; 1991). However, apart from these general issues, my main interest lies in more specific studies of acoustic phonetics. In recent years, these have been extended to include prosody. References to work carried out at the KTH may be found in the major review articles listed below.

References:

Fant, G.(1990a): "Speech research in perspective," *Speech Communication* **9**, pp. 171-176.

Fant, G.(1990b): "The speech code. Segmental and prosodic features," pp. 1389-1397 in *Proc. Int. Conf. on Spoken Language Processing, Kobe, Japan, Vol. 2,* Acoust.Soc. of Japan, Tokyo.

Fant, G.(1991): "What can basic research comcopntribute to speech synthesis?" to be publ. in *J. of Phonetics.*