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

U.S. - Japan joint seminar on dynamic aspects of speech production: 'Key-note speech'

Fant, G.

KTH Computer Science and Communication

journal: STL-QPSR
volume: 17
number: 4
year: 1976
pages: 021-027

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

U. S. - Japan Joint Seminar on Dynamic Aspects of Speech Production, December 7-10, 1976, Tokyo, Japan

"Key-Note Speech" by G. Fant

When collecting my thoughts on the topic of this seminar, I am aware of that I have spent more time in life worrying about the static aspects than about the dynamic aspects of speech production. What I have to say is accordingly the result of an effort to keep up with an area of research in which you are the experts. I have learned much from the impressive collection of papers you have submitted and they have given me the impetus for reorganizing my thinking on the subject.

We cannot understand the nature of speech without being aware of the dynamics. To quote Stetson - "Speech is movements made audible". The success of future work in speech synthesis by rule, in automatic speech recognition, in advancing the basis of general phonetics, and in the many applications in improved teaching methods in foreign languages, in deaf school teaching, in speech pathology, and in aids for handicapped will depend on how well we understand the dynamics of speech production.

In this broader frame of applications we are really concerned with the overall speech code, dynamic as well as static, on the levels of both production, speech wave, and perception. When studying speech production, it is close at hand to pose questions about speech wave correlates and perception just as when we attempt to study the perception of speech wave patterns and seek a guidance in production patterns.

A study of the dynamics of speech production presumes a knowledge of the static aspects, that is, the dimensionality of possible states of the production system. Secondly, we would like to know the general constraints on the dynamics of possible sequences and combinations of states and the role of feedback control mechanisms. Given this insight we are then prepared to describe the speech code, how a specific message in a specific language or dialect spoken by a specific type of speaker in a specific mode will be realized. It is obvious that we have a long way to go before we have come that far and can fill the existing gaps between linguistics and phonetics. The rules of generative phonology do not include the transforms for predicting the dynamics at speech production and are thus incapable of truly describing phonetic events.

The search for the speech code thus implies coordinated studies of both speech production, speech wave, and speech perception. What is hidden from insight in attempts to study production may be inferred from the speech wave. As we all know, the speech wave provides us with a complete record of the spoken message but it is a most complex task to sort out the detailed fine structure and the interrelations between measured parameters in search for a reliable representation of underlying production events.

Vocal tract line analogue models are now being developed to allow a reasonable accurate synthesis specially of vowels. These models still need to be improved with respect to consonants and with respect to the choice of control parameters. Also our insight in the anatomical differences between male and female vocal tracts is not very deep. Almost all of our synthesizers are scaled after a male prototype.

It would be a considerable gain in experimental techniques for speech production studies if it were possible to extract reliable vocal tract area function data from a speech wave processing. An impressive amount of theoretical work in this direction has been developed during the last year but I do not feel that we have reached a sufficiently high level of accuracy. Some of the seminar papers are devoted to this topic and to the derivation of underlying articulatory movements and commands. We are still faced with problems of uniqueness and of overidealizations of the actual vocal tract transmission properties.

So for the time being we have taken resource to x-ray studies of the vocal tract. Such data are badly needed but we hope that the low dosage computer controlled microbeam system developed by Fujimura and others will provide a break-through. On the other hand, we already have some interesting material collected by means of conventional x-ray techniques. I especially have in mind the studies undertaken by Kenneth Moll and his associates at Iowa University.

There exists a great variety of mechanical, acoustical, optical, electrical, and magnetic field methods for direct observations and it will be especially interesting to see how they may supplement other methods. The dynamic palatography developed at the Research Institute of Logopedics and Phoniatrics has provided us with interesting observations,

e. g., on differences in articulation of tense and lax stops. The study of laryngeal articulations and of the dynamics of glottal adjustments has been greatly enhanced by the fiberoptic methods introduced by Sawashima and others. These techniques are also useful for studies of the velopharyngeal opening during speech.

It is evident that our insight in the speech encoding and decoding stages becomes more limited the further away from the speech wave a particular stage is located. In order to gain a maximum of information from studies of speech production, it is apparently valuable to combine as many simultaneous measurements as possible, e. g., to supplement EMG records with a tracking of articulatory movements and resulting speech wave characteristics. As we all know, EMG data can be misleading since the role of a single muscle in a coordinated activity of many muscles may vary with speakers and context and one is not always sure to hit the right muscle at the right place to ensure a representative activity. It is therefore not advisable to draw too far reaching conclusions from EMG data alone.

In general, we are faced with problems of interpreting fragmentary or at the best incomplete data to supply evidence for our modelling of production processes. It is like driving a long nail into a computer with the hope that it may be used as a pick-up electrode to reveal the architecture of the computer, its programming language and operational system. No wonder neurophysiologists sometimes complain that we speech scientists are too hasty in making far reaching conclusions about neural correlates to linguistic units. A greater overall insight in the system is needed, as pointed out by Moll.

The general theory of speech analysis provides a frame for our modelling of production. I shall recall some stages in the development of speech wave models.

The naive view of speech is that we produce a sequence of discrete successive sounds each corresponding to a letter or to a combination of letters in the alphabet. With a less naive definition of message units, this basic notion amounts to an expectancy of finding one quasi-stationary speech segment for each phoneme. Although this is almost the case for many fricatives, the general situation is, as we all know, a many-to-one relation in either direction from phonemes to speech wave events.

Acoustic segmentation may involve arbitrary decisions. To quote Stevens "The truth about segmentation is that you cannot".

Instead of discrete pattern breaks we are often faced with continuous variations. Another extreme formulation of the dilemma of speech analysis originates from Charles Hockett in 1958: "Imagine the phonemes as a succession of eggs ordered in a row. Next crush them against a moving belt and the resultant mess is the speech wave." There are, however, interesting attempts at present to provide order into this mess by analysis-by-synthesis methods to derive underlying articulatory commands. I am thinking of the direct acoustic approach of Fujisaki and the articulatory modelling of Nakajima and of Shirai.

On the other hand, one can view the complexity of speech encoding from a more purposeful view to bring out the efficiency. Liberman has pointed out that the decomposition of a message into a superposition of elementary phonetic events in space and time matches the information processing capacity of the human brain. Parallel processing and combinations of short and long time memory function ensure a high speed of encoding.

Returning to the history of speech analysis, the second generation model implies that speech is an alternation between quasi-stationary and transitional events. Consonants are identified by a combination of these. This picture grew out of the early pioneering work at Haskins Laboratories. It has inspired strategies of speech synthesis by rule and has been most influential in general phonetic theory. It is interesting to note that only very recently the importance of catching cues from transitional events has been fully recognized by people working in speech recognition.

Our present more complete models allow for a continuity of movements subject to rules of contextual reorganizations, coarticulations, and reductions eliminating the sharp distinction between transitional and more continuous events. On the other hand, we should be aware of the pronounced non-uniformities in the temporal spread of information bearing cues to revise some of the basic rules of perceptual significance. The base rule stating that stationary segments signal the manner and transitions signal the place of articulation has more exceptions than we might expect. Thus, the nasal murmur of [m], [n], and [ŋ] may contain

strong place cues and the first part of the vowel after the release of a consonant is generally as important as the preceding stationary part in signalling manner cues, e.g. a lateral versus a nasal. The first 20 milliseconds after release of a [t] or [p] may be more important than the following transition of the F-pattern, whereas the transitions carry important information about vowel identity.

One main point of this excursion is that the durations and time constants involved in the dynamics of speech production do not always have simple counterparts in perception and that this is one of the many arguments for pursuing further work on the dynamics of speech perception.

In spite of our insight in the highly dynamic characteristics of speech, we still have a choice of phrasing our concepts of invariants in more or less static terminology. When discussing discrete message units and articulatory targets, we may, as Ken Stevens, restrict ourselves to ensembles of possible discrete states, but if we want to describe the speech events and the mechanisms of perception, we must tackle the dynamic aspects. Have the static aspects larger psychological reality than the dynamic aspects?

This is a kind of challenge to push further into the still very much unknown territory of perception or to retreat back to a study of production models and be content with a view that perception is just the inverse of production: a search for equivalent motor commands. A clever strategist, and I know Frank Cooper is one, would probably take an intermediate view to make a temporary retreat into the domains of production in order to reorganize his troupes and make a more powerful charge into the domains of perception.

What do we know about the dynamics of speech production? The history is well covered in our contributed papers and I shall therefore not attempt a duplication. We appear to agree on the following main issue:

Speech is a feedback mediated, output-oriented integration of movements in space and time executed by a complex of exhibitory and inhibitory muscle activities.

An essential point is that auditory, tactile, and proprioceptive feedback loops operate to ensure the most adequate output independent of

disturbances and of the immediate contextual frame of preceding and following states.

I could quote Roman Jakobson: "We speak to be heard in order to be understood."

Exactly what aspects of output invariance are preserved is a matter of further studies. One characteristic that has been observed in early studies of Öhman and recently by the research group at Iowa University is the tendency toward constant duration of consonant-vowel transitions independent of context.

The nature and timing of right to left carry-over effects and left to right anticipatory coarticulation should be studied in greater detail. When do we allow an articulator to start moving towards a following target? Is it when it is free to move and does not impede the overall pattern? If so, how early? In short, how great is the asymmetry of coarticulation in various contexts and to what extent are articulatory components traditionally assigned to one phonéme segment synchronized. What are the general rules for synchronizing phonatory and articulatory states?

There are a number of problems, largely concerning prosodics, which are of interest to consider in this connection but which have not been included in the background material. I have in mind principles of redistribution of articulatory energy with changes in place of emphasis and stress and the relation of such phenomena to the tense-lax opposition. We do not have any material in this seminar on positional and contextual dependencies of vowel and consonant durations. Björn Lindblom has some new ideas about the structuring of durations that I might review in one of our sessions.

In rounding up this overview I realize that the study of speech production invites to many opposite views and paradoxal statements: speech is discrete and continuous; fast and slow; sloppy and precise; simple and complex. It is not the question of one or the opposite but <u>both</u> so we miss the perspective if we set out to proof that only one of the two alternatives is true or untrue. The same holds for issues such as segmentability and whether features, phonemes or syllables are to be considered as minimal units.

In conclusion we are much better off if we set our goal to describe speech more fundamentally descriptive with respects to actual mechanisms and performance than to prove any extreme view.

To speak is so simple - a child can do what 20 wise men (and three women) at this seminar do not quite understand. That is part of the charms of this ever challenging research area.